Université de Toulouse

# THÈSE

**En vue de l'obtention du**

## DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

**Délivré par :** *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

---

**Présentée et soutenue le** *20 décembre 2017* **par :**

**François-Xavier Decroix**

---

# Apprentissage en ligne de signatures audiovisuelles pour la reconnaissance et le suivi de personnes au sein d'un réseau de capteurs ambiants

---

**JURY**

| | | |
|---|---|---|
| Serge MIGUET | Université Lumière Lyon 2 | Président du Jury |
| Patrick LAMBERT | Université Savoie Mont Blanc | Rapporteur |
| Michel VACHER | Université Grenoble Alpes | Rapporteur |
| Isabelle FERRANE | Université Toulouse III | Examinatrice |
| Frédéric LERASLE | Université Toulouse III | Directeur de thèse |
| Julien PINQUIER | Université Toulouse III | Co-directeur de thèse |

---

**École doctorale et spécialité :**
*MITT : Signal, Image, Acoustique et Optimisation*

**Unité de Recherche :**
*Institut de Recherche en Informatique de Toulouse (UMR5505)*

**Directeur(s) de Thèse :**
*Frédéric LERASLE* et *Julien PINQUIER*

**Rapporteurs :**
*Patrick LAMBERT* et *Michel VACHER*

À Anne Kieffer et Simone Pichelin

# Remerciements

Je voudrais tout d'abord remercier chaleureusement mes directeurs de thèse, Frédéric Lerasle et Julien Pinquier, ainsi qu'Isabelle Ferrané, pour leur très grande patience, leur implication et la liberté qu'ils m'ont accordé pendant ces 4 ans. Ces travaux n'auraient pas été possibles sans leurs conseils, leurs encouragements et les nombreuses relectures à toute heure du jour et de la nuit et je les en remercie.

Je souhaiterais ensuite remercier les rapporteurs de cette thèse, Patrick Lambert et Michel Vacher pour leur retours constructifs, leur prévenance et leur ouverture face au caractère novateur de ces travaux, en particulier pendant la soutenance. Je remercie également Serge Miguet d'avoir accepté d'être examinateur et pour sa bienveillance lors de la soutenance de cette thèse.

Un grand merci également aux membres de l'équipe SAMoVA qui me supportent depuis presque cinq ans et mon stage de fin d'études. Une mention spéciale à Christine pour avoir partagé son bureau avec moi et pour ses anecdotes inénarrables. Je salue aussi l'équipe RAP du LAAS et tous les doctorants de la team ICU, en particulier Ali et Christophe qui ont été d'une grande aide et un soutien précieux dans le bureau et au quotidien.

Même s'il m'est malheureusement impossible d'être exhaustif, je ne remercierai jamais assez les amis de l'ENSEEIHT qui ont toujours répondu présent dans les périodes difficiles : Mathieu, Jérémy, Dominique, Martin pour avoir vécu ça ensemble, et bien évidemment Pedro, Léo et Ivan pour l'incroyable aventure humaine et musicale de ces sept dernières années.

Enfin, je ne pourrai conclure ces remerciements sans me tourner vers ma famille et en particulier mes parents qui, malgré la distance, ont inlassablement cru en moi tout au long de mes études. Leur soutien infaillible a été indispensable à la réalisation de ces travaux.

# Résumé

L'opération neOCampus, initiée en 2013 par l'Université Paul Sabatier, a pour objectif de créer un campus connecté, innovant, intelligent et durable en exploitant les compétences de 11 laboratoires et de plusieurs partenaires industriels. Pluridisciplinaires, ces compétences sont croisées dans le but d'améliorer le confort au quotidien des usagers du campus (étudiants, corps enseignant, personnel administratif) et de diminuer son empreinte écologique. L'intelligence que nous souhaitons apporter au Campus du futur exige de fournir à ses bâtiments une perception de son activité interne. En effet, l'optimisation des ressources énergétiques nécessite une caractérisation des activités des usagers afin que le bâtiment puisse s'y adapter automatiquement. L'activité humaine étant sujet à plusieurs niveaux d'interprétation nos travaux se focalisent sur l'extraction des déplacements des personnes présentes, sa composante la plus élémentaire.

La caractérisation de l'activité des usagers, en termes de déplacements, exploite des données extraites de caméras et de microphones disséminés dans une pièce, ces derniers formant ainsi un réseau épars de capteurs hétérogènes. Nous cherchons alors à extraire de ces données une signature audiovisuelle et une localisation grossière des personnes transitant dans ce réseau de capteurs. Tout en préservant la vie privée de l'individu, la signature doit être discriminante, afin de distinguer les personnes entre elles, et compacte, afin d'optimiser les temps de traitement et permettre au bâtiment de s'auto-adapter. Eu égard à ces contraintes, les caractéristiques que nous modélisons sont le timbre de la voix du locuteur, et son apparence vestimentaire en termes de distribution colorimétrique.

Les contributions scientifiques de ces travaux s'inscrivent ainsi au croisement des communautés parole et vision, en introduisant des méthodes de fusion de signatures sonores et visuelles d'individus. Pour réaliser cette fusion, des nouveaux indices de localisation de source sonore ainsi qu'une adaptation audiovisuelle d'une méthode de suivi multi-cibles ont été introduits, représentant les contributions principales de ces travaux. Le mémoire est structuré en 4 chapitres. Le premier présente un état de l'art sur les problèmes de ré-identification visuelle de personnes et de reconnaissance de locuteurs. Les modalités sonores et visuelles ne présentant aucune corrélation, deux signatures, une vidéo et une audio sont générées séparément, à l'aide de méthodes préexistantes de la littérature. Le détail de la génération de ces signatures est l'objet du chapitre 2. La fusion de ces signatures est alors traitée comme un problème de mise en correspondance d'observations audio et vidéo, dont les détections correspondantes sont cohérentes et compatibles spatialement, et pour lesquelles deux nouvelles stratégies d'association sont introduites au chapitre 3. La cohérence spatio-temporelle des observations sonores et visuelles est ensuite traitée dans le chapitre 4, dans un contexte de suivi multi-cibles.

# Abstract

The neOCampus operation, started in 2013 by Paul Sabatier University in Toulouse, aims to create a connected, innovative, intelligent and sustainable campus, by exploiting the skills of 11 laboratories and several industrial partners. These multidisciplinary skills are combined in order to improve users (students, teachers, administrative staff) daily comfort and to reduce the ecological footprint of the campus. The intelligence we want to bring to the campus of the future requires to provide to its buildings a perception of its intern activity. Indeed, optimizing the energy resources needs a characterization of the user's activities so that the building can automatically adapt itself to it. Human activity being open to multiple levels of interpretation, our work is focused on extracting people trajectories, its more elementary component. Characterizing users activities, in terms of movement, uses data extracted from cameras and microphones distributed in a room, forming a sparse network of heterogeneous sensors. From these data, we then seek to extract audiovisual signatures and rough localizations of the people transiting through this network of sensors. While protecting person privacy, signatures must be discriminative, to distinguish a person from another one, and compact, to optimize computational costs and enables the building to adapt itself. Having regard to these constraints, the characteristics we model are the speaker's timbre, and his appearance, in terms of colorimetric distribution. The scientific contributions of this thesis are thus at the intersection of the fields of speech processing and computer vision, by introducing new methods of fusing audio and visual signatures of individuals. To achieve this fusion, new sound source location indices as well as an audiovisual adaptation of a multi-target tracking method were introduced, representing the main contributions of this work. The thesis is structured in 4 chapters, and the first one presents the state of the art on visual reidentification of persons and speaker recognition. Acoustic and visual modalities are not correlated, so two signatures are separately computed, one for video and one for audio, using existing methods in the literature. After a first chapter dedicated to the state of the art in re-identification and speaker recognition methods, the details of the computation of the signatures is explored in chapter 2. The fusion of the signatures is then dealt as a problem of matching between audio and video observations, whose corresponding detections are spatially coherent and compatible. Two novel association strategies are introduced in chapter 3. Spatio-temporal coherence of the bimodal observations is then discussed in chapter 4, in a context of multi-target tracking.

# Liste des symboles

**ACC** Analyse Canonique des Corrélations

**ACF** Aggregate Channel Features

**ACP** Analyse en Composantes Principales

**BD** Base de Données

**CMC** Cumulative Matching Curve

**DCT** Discrete Cosine Transform

**DET** Detection Error Tradeoff

**EM** Expectation Maximisation

**ER** Error Rate

**FFT** Fast Fourier Transform

**GMM** Gaussian Mixture Model

**HOG** Histograms of Oriented Gradients

**HSV** Hue Saturation Value

**ICAR** Incorrect to Correct Associations Ratio

**ID** Identité

**IPAV** Indice de Proximité Audio Vidéo

**JFA** Joint Factor Analysis

**LLR** Log Likelihood Ratio

**MAE** Mean Absolute Error

**MAP** Maximum A Posteriori

**MCMCDA** Markov Chain Monte Carlo Data Association

**MCMC** Markov Chain Monte Carlo

**MFCC** Mel Frequency Cesptral Coefficients

**MLE** Maximum Likelihood Estimation

**MOTA** Multiple Object Tracking Accuracy

**MOTP** Multiple Object Tracking Precision

**MOT** Multiple Object Tracking

**MSE** Mean Squared Error

**NCA** Normalized Correct Associations

**PoI** Point of Interest

**RGB** Red Green Blue

**RoI** Region of Interest

**SDALF** Symetry Driven Accumulation of Local Features

**SNR** Signal to Noise Ratio

**SRMR** Speech to Reverberation Modulation energy Ratio

**UBM** Universal Background Model

**VAD** Vocal Activity Detection

**ZCR** Zero Crossing Rate

# Table des matières

# Introduction

## Contexte, enjeux et problématique

### La ville de demain, les enjeux d'une transition éco-responsable

La transition d'une économie fondée sur l'agriculture à une économie fondée sur l'industrie de masse, les services et les nouvelles technologies est responsable d'une explosion de la taille des villes dans le monde et d'un exode rural massif. Le taux de la population urbaine mondiale était estimé à un tiers de la population totale en 1960. Il s'est élevé à 54% en 2014 et devrait atteindre 66% en 2050 selon un rapport des Nations Unies [UN14]. Cette urbanisation massive pose alors de nouveaux enjeux écologiques : les villes sont responsables à 80% des émissions de gaz à effet de serre, provoquant l'urgence d'une transition énergétique globale (COP 21), mais aussi locale. Le concept de ville intelligente (« *Smart City* ») apparaît dès les années 80 dans les mégalopoles asiatiques, comme Singapour ou Shanghai et se popularise dans de nombreuses grandes villes. Il est fondé sur une optimisation des coûts, de l'organisation et du bien-être des habitants. La ville de Toulouse présente ainsi un plan d'investissement public de 500 millions d'euros en 2020 pour transformer Toulouse en « Open Métropole » [1] sur des objectifs de mobilité, de respirabilité et de sûreté. Des initiatives apparaissent également au niveau universitaire, citons parmi celles-ci le Smart Campus de l'Université Versailles-Saint-Quentin [2], l'Open Smart Campus de l'Université Joseph Fourier de Grenoble-Alpes [3], ou encore le programme Living Smart Campus de l'Université de Twente [4], aux Pays-Bas. L'opération neOCampus s'inscrit dans cette démarche.

### Une mise à l'échelle : l'opération neOCampus

Initiée en 2013 par Bertrand Monthubert, président de l'Université Paul Sabatier de Toulouse, l'opération neOCampus [5] a pour objectif de faire un campus innovant, intelligent et durable, transférant à son échelle les concepts de la ville intelligente. À ce jour, le savoir-faire de plus de 10 laboratoires est mis à contribution pour réduire son empreinte écologique et améliorer le confort des usagers, étudiants, enseignants et personnels administratifs. Les axes de recherche de ce projet sont ainsi pluridisciplinaires et croisent plusieurs communautés scientifiques (énergie, numérique, génie électrique, matériaux, robotique, biodiversité...). Le socle commun est la collecte et l'exploitation de données dans une approche « Open Data » participative ainsi qu'un pilotage distribué. Ces données sont extraites d'un réseau de capteurs hétérogènes disséminés sur

1. http://www.toulouse-metropole.fr/projets/smart-city
2. http://www.smartgrids-cre.fr/index.php?p=smart-campus
3. http://www.filiere-3e.fr/2014/06/25/open-smart-campus-etudiants-grenoblois-inventent-campus-demain/
4. https://www.utwente.nl/en/organization/news-agenda/special/2016/living-smart-campus/
5. https://www.irit.fr/neocampus/fr/

le campus de manière éparse : s'il peut communiquer avec les autres composantes du réseau, un capteur doit pouvoir fonctionner en autonomie, et ne pas être contraint par la topologie générale du réseau, dans une philosophie « plug and play ». Nous désignons par bâtiment intelligent un bâtiment capable non seulement d'agréger les informations issues des capteurs mais également de s'adapter à l'activité de ses usagers en fonction des informations collectées. Le bâtiment ADREAM du LAAS-CNRS [6], inauguré en 2012, en est un exemple par sa gestion énergétique optimisée en fonction des besoins des différents utilisateurs. Au niveau du campus, 3 salles d'enseignement de l'Université ont déjà été ainsi équipées de capteurs ambiants de différents types (luminosité, présence, caméras, microphones,...), et font figure de démonstrateurs des outils développés dans le cadre de neOCampus : gestion automatique des stores, monitoring des ressources énergétiques... Une des salles d'expérimentation est illustrée sur la figure 1. Cette thèse ayant démarré en même temps que le projet neOCampus, il était primordial dans ce cadre applicatif très large et ambitieux de s'intéresser à des fonctionnalités bas niveau (ou atomiques) de ce réseau large échelle de capteurs intelligents et communicants. Les travaux de cette thèse se focalisent donc sur le monitoring d'activité à partir de caméras et de microphones, et le cadre applicatif choisi est une salle pédagogique, cœur des activités du campus.

FIGURE 1 – Photographie d'une salle d'enseignement de l'Université Paul Sabatier équipée de capteurs.

### Objectifs de nos travaux : focus sur l'apprentissage de signatures audiovisuelles

À l'inverse d'autres modalités, plus neutres, le monitoring audiovisuel pose de réelles questions éthiques et juridiques. Loin de la surveillance et du contrôle de masse par l'identification biométrique, thème cher à la littérature d'anticipation, il vise une compréhension non nominative de l'activité des usagers, dont les interactions principales sont visuelles et sonores. L'objectif de ces travaux de thèse est alors d'extraire des flux sonores et visuels des capteurs ambiants une information non intrusive pour les usagers qui puisse aider à la compréhension de leur activité et de leurs interactions, entre eux ou avec les infrastructures présentes. L'activité pouvant endosser plusieurs définitions, selon les domaines d'étude, il est important de rester à une description bas niveau des percepts audiovisuels. Pour monitorer l'activité des usagers à l'échelle d'un bâtiment,

6. `https://www.laas.fr/public/fr/le-projet-adream`

il faut ainsi ré-identifier ces usagers à partir du réseau épars instrumentant cet environnement et ainsi apprendre en ligne, leurs signatures audiovisuelles non-nominatives.

Nous entendons par **signature** une caractérisation à fort pouvoir discriminant de chaque usager cible. Celle-ci ne porte aucune information personnelle et se place au-delà de la biométrie (reconnaissance faciale, puces RFID...) : une identité audiovisuelle virtuelle apprise en ligne est assignée à chaque individu détecté dans la zone couverte par les capteurs. Cette ré-identification directe est effectuée en quasi temps réel et sur une échelle temporelle limitée à quelques heures, nous permettant ainsi d'exploiter l'apparence vestimentaire des usagers pour apprendre notre signature vidéo. Une carte des déplacements (donc des activités) et des interactions audiovisuelles peut ainsi être établie et transmise aux actionneurs et superviseur du bâtiment intelligent.

### Contributions et verrous scientifiques

Les objectifs présentés ci-dessus placent ces travaux au croisement d'une communauté vision et d'une communauté audio qui ne partagent pas toujours les mêmes codes. Un monitoring audio s'affranchit généralement de la localisation de la source sonore, à l'inverse du monitoring vidéo. La difficulté majeure à laquelle se sont confrontés ces travaux est le caractère novateur de la problématique, notamment dû à la faible densité et l'hétérogénéité des capteurs, qui la met en marge des problématiques classiques de fusion audiovisuelle.

**Association des modalités audio et vidéo à travers un réseau épars de capteurs.** Les représentations sonores et visuelles d'individus sont des problématiques largement étudiées : l'état de l'art peut être considéré suffisamment riche pour notre application. En revanche la fusion de ces modalités à travers un réseau épars de capteurs hétérogènes représente un défi important. Sans identification réelle, il est impossible de lier une signature sonore et une signature visuelle en s'appuyant uniquement sur leur caractérisation, les deux modalités étant décorrélées (une voix peut correspondre à n'importe quelle apparence, et inversement). Ce « verrouillage » audiovisuel exploitera la cohérence spatio-temporelle des percepts, particulièrement difficile à observer dans un réseau épars de capteurs. En effet ce verrouillage est sensible à l'encombrement de la scène observée, à l'intermittence des percepts et aux erreurs pouvant intervenir à chaque composante des systèmes de génération des signatures.

**Localisation monocapteur des sources sonores.** Une tâche inhérente à la fusion évoquée ci-dessus est la localisation des sources. Si c'est une problématique résolue en vision, par la calibration des caméras, elle est bien plus ardue pour les sources sonores. De manière analogue à la perception humaine, la localisation précise d'une source sonore repose sur l'exploitation d'indices multi-auraux, condition hors de notre contexte épars et « plug and play ». Chaque percept audio monophonique doit pouvoir être associé indépendamment à une estimation spatiale de la source, sortant ainsi des contextes courants des problématiques de fusion audiovisuelle dans la littérature. En effet celles-ci exploitent généralement des chaînes de microphones pour l'estimation d'un azimut de la source sonore et un tel équipement sort de notre contexte applicatif.

**Intégration des signatures dans un suivi multimodal.** L'établissement de la carte des déplacements des usagers étant l'un des objectifs de ces travaux, une étape de suivi est nécessaire. L'intégration des signatures sonores et visuelles des usagers dans un traqueur pour un suivi multimodal représente alors la dernière contribution de ces travaux.

## Organisation du manuscrit

Le mémoire est structuré en 4 chapitres.

Le chapitre 1 expose différentes méthodes de la littérature dans le but d'identifier les tendances en terme de génération de signatures sonores et visuelles d'individus. Ceci nous permet de choisir les outils, techniques, bases de données et métriques pertinentes, eu égard à notre contexte applicatif. Concernant la reconnaissance de locuteurs, il s'agit de répondre à la question : « qui parle, parmi les locuteurs connus ? ». Pour la ré-identification visuelle de personnes, la question est : « à quelle identité locale associer l'individu détecté ? ». Ce chapitre vise ainsi à positionner notre démarche et à exposer les concepts et le cadre d'étude adoptés par la suite.

Les méthodes choisies pour la ré-identification visuelle de personnes ainsi que la reconnaissance de locuteurs sont détaillées dans le chapitre 2 et évaluées sur une plate-forme expérimentale personnelle, face à l'absence de base de données publiques correspondant précisément à notre contexte. Les problématiques les plus proches, notamment l'identification audiovisuelle du locuteur actif, ne se placent en effet pas dans un contexte ambiant mais dans des configurations plus contrôlées, e.g. de visio-conférence. Nous avons alors acquis un jeu de données audiovisuelles avec un petit nombre de participants évoluant librement dans l'espace couvert par le champ des capteurs. L'enjeu de cette tâche est de valider l'adaptabilité des méthodes pour une définition robuste des signatures dans notre contexte.

Le chapitre 3 présente deux méthodes de fusion de signatures audiovisuelles dans un cas mono-cible, à travers l'exploration d'indices de localisation des sources sonores. Face à l'impossibilité d'une estimation de la position en deux dimensions du locuteur actif, nous extrayons une mesure de distance par l'exploitation d'une propriété acoustique de la pièce, sa réverbération. Cette mesure permet la définition de zones de saillance audiovisuelle dans lesquelles peuvent être fusionnées les signatures audio et vidéo d'un individu. Dans un second temps nous étudions le mélange de plusieurs mesures pour améliorer l'estimation de cette distance. Ce chapitre résout ainsi le problème du verrouillage audiovisuel des signatures en contexte mono-cible.

Le chapitre 4 ouvre au suivi multi-cibles. Nous adaptons une approche existante de suivi par l'utilisation des signatures sonores et visuelles des individus. Celles-ci sont intégrées séparément et fusionnées au sein de l'outil de suivi. En effet, le verrouillage des identités audio et vidéo étant limité à des scenarii mono-cibles ou dans des environnements peu encombrés hors suivi, l'analyse de la cohérence et compatibilité spatio-temporelle des percepts audio et vidéo permet de renforcer cette fusion dans le cas multi-cibles.

# Chapitre 1

# Problématique et positionnement des travaux

**Sommaire**



## Introduction

Ce chapitre détaille notre problématique générale puis présente les méthodes et outils existants qui traitent de génération de signatures sonores et visuelles d'individus. Celles-ci se réfèrent aux tâches de ré-identification pour la signature visuelle et de reconnaissance de locuteurs pour la signature sonore. Un état de l'art exhaustif sort du cadre de ce chapitre, nous énumérons ici les principaux outils, techniques, bases de données et métriques d'évaluation afin de positionner nos

travaux et justifier, autant que possible, nos choix eu égard à cette littérature et notre contexte applicatif.

La section 1.1 rappelle la problématique et précise le cahier des charges associé. La section 1.2 (respectivement 1.3) énumère les principales méthodes de reconnaissance de locuteurs (respectivement de ré-identification de personnes). Nous relevons, au fur et à mesure, les outils, techniques, etc. pertinents pour notre contexte applicatif.

## 1.1 Problématique générale de la thèse

Précisons ci-après la problématique, les verrous et hypothèses sous-jacentes associées, afin de positionner nos investigations eu égard à la littérature.

### 1.1.1 Cahier des charges

**Réseau épars de capteurs hétérogènes.** Un verrou majeur réside dans l'instrumentalisation de la plate-forme expérimentale. Nous privilégions des capteurs bas coûts pour limiter les frais de déploiement large échelle (à terme) du réseau. Nous utiliserons alors des caméras et microphones standards du commerce, installés de manière éparse, et dont les champs sont disjoints ou partiellement joints. Notre focus étant sur l'association audio-vidéo, nous nous concentrons alors sur l'exploitation d'un couple caméra/microphone.

**Traitement des données en environnement fermé.** Une hypothèse simplificatrice, mais courante sur cette problématique de ré-identification de personnes, est celle du « *closed set* » , soit un environnement fermé, ici non seulement spatialement mais également temporellement : la ré-identification s'effectue sur des sessions à durée réduite (entre 30 minutes et deux heures). Nous considérons ainsi que les cibles perçues/détectées « en live » font partie d'une base de données de cibles préalablement apprises en début de session. En pratique, nous supposons que tous les accès au bâtiment sont instrumentés de capteurs afin de générer une première perception des cibles et donc incrémenter la base de données.

**Inférence quasi temps réelle à horizon temporel borné.** Notre application de bâtiments auto-adaptatifs nécessite que le monitoring s'effectue en « live » et que les traitements n'excèdent pas quelques minutes. De plus, les sessions d'acquisition et le traitement des flux sensoriels sont bornés temporellement, de l'ordre de quelques heures maximum. Ceci permet de monitorer les déplacements des usagers sur une journée.

**Déploiement peu contraignant du réseau multi-capteurs.** Les capteurs sont gérés en mode « plug and play ». Les phases d'étalonnage géométriques intra- ou inter-capteurs et temporelles (synchronisation inter-capteurs) doivent induire une intervention humaine minimale lors de leur installation/retrait.

**Intermittence des modalités.** Les percepts audio et vidéo sont par nature intermittents. Particulièrement en audio, où les discours ne se superposent que rarement dans le contexte qui nous intéresse, mais également en vision où des personnes cibles, en mouvement, peuvent être temporairement occultées par du mobilier ou d'autres personnes. La perception du (ou des) cible(s), à chaque instant, n'est donc pas systématiquement multimodale !

### 1.1.2 Synthèse

Eu égard à ces spécifications, les signatures doivent être discriminantes et robustes sur un horizon de quelques heures, tout en fournissant des inférences de ré-identification quasi temps réel, sans contrainte forte sur la configuration multi-capteurs et avec de possibles intermittences sur les observations associées. Enfin, nous supposons que l'environnement est fermé, c'est-à-dire composé d'une base de données de cibles (à ré-identifier) préalablement établie.

Nous procédons maintenant à la présentation d'un état de l'art non exhaustif en audio puis en vidéo afin d'exhiber des signatures sonore et visuelle compatibles avec notre application.

## 1.2 Reconnaissance de locuteur

Nous donnons d'abord un positionnement de nos travaux vis à vis de cette modalité de reconnaissance de locuteurs, ainsi qu'une description des méthodes usuelles. Pour finir, une mise en œuvre est effectuée via les outils et les cadres d'évaluations existants.

### 1.2.1 Positionnement de nos travaux

Nous nous intéressons ici aux interactions impliquant potentiellement une activité de parole car porteuse d'information pouvant caractériser la personne qui parle, ou plutôt sa voix. Parole et langage se distinguent depuis le début des travaux en linguistique [Sau16] et nous nous intéressons ici à la parole en tant que signal physique, produit par un individu et capté par les microphones avec éventuellement une dégradation due au bruit ambiant et/ou à la qualité du capteur lui même ou à la distance locuteur microphone. Nous ne nous intéresserons pas au contenu du message qui transite du locuteur vers ses éventuels interlocuteurs. Nous nous plaçons dans un cadre « indépendant de ce qui est dit, donc indépendant du texte » sans analyse linguistique et sans utilisation des informations qu'elle pourrait dégager (nom des personnes, et autres infos personnelles, etc.).

Les travaux de cette thèse se focalisent sur la reconnaissance de locuteurs, soit l'identification de la singularité de chaque individu en terme de production vocale. Nous présentons par la suite les différentes tâches et méthodes de la littérature relatives à cette problématique.

#### 1.2.1.1 Tâches de reconnaissance liées au locuteur

En fonction de l'application ciblée, la reconnaissance du locuteur peut s'exprimer comme une tâche de vérification, d'identification ou de structuration (segmentation et regroupement ou « *speaker diarization* » en anglais). Ces tâches sont illustrées dans le tableau 1.1.

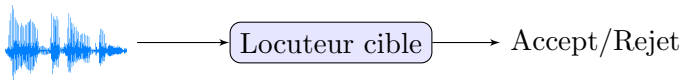
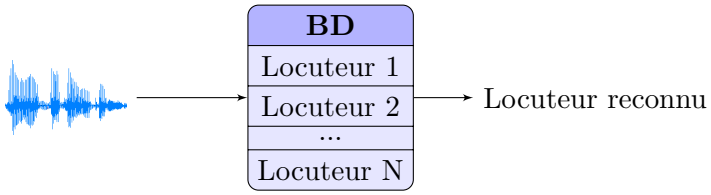
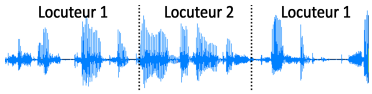
- **Vérification du locuteur :** nous testons l'hypothèse que le signal de parole en entrée du système provient d'un locuteur cible, dont les caractéristiques sont préalablement apprises,
- **Identification du locuteurs :** l'émetteur du signal inconnu est recherché parmi un ensemble de $N$ locuteurs, préalablement appris,
- **Structuration en locuteurs :** un signal audio est découpé à chaque changement de locuteur, et les segments ainsi produits sont agrégés par locuteur.

Les tâches de vérification et d'identification du locuteur se distinguent uniquement par leur bloc décisionnel. Dans les deux cas, le système calcule un score de confiance dans l'hypothèse que

le segment de parole inconnu soit produit par un locuteur préalablement appris. En vérification de locuteur, ce score est confronté à un seuil $\theta$ pour l'acceptation ou le rejet de l'hypothèse. En identification, $N$ hypothèses sont testées, une par locuteur, et le choix du locuteur est défini par le meilleur des $N$ scores correspondant à chaque locuteur de la Base de Données (BD).

Dans la suite, nous nous concentrons uniquement sur ces deux tâches sous le même terme de « reconnaissance du locuteur ».

TABLE 1.1 – Illustration des tâches de vérification, d'identification et de structuration en locuteurs.

| | |
|---|---|
| Vérification de locuteur |   |
| Identification de locuteurs |   |
| Structuration en locuteurs |   |

### 1.2.2 Architecture des systèmes de reconnaissance de locuteurs et méthodes associées

Cette partie détaille l'architecture traditionnelle sur laquelle sont construits les systèmes de reconnaissance de locuteurs. Si la littérature est fournie et dynamique sur le sujet, les approches s'appuient sur une même séquence (figure 1.1). Celle-ci est composée de trois étapes principales : l'extraction de paramètres propre à chaque locuteur, la modélisation du locuteur et la classification.

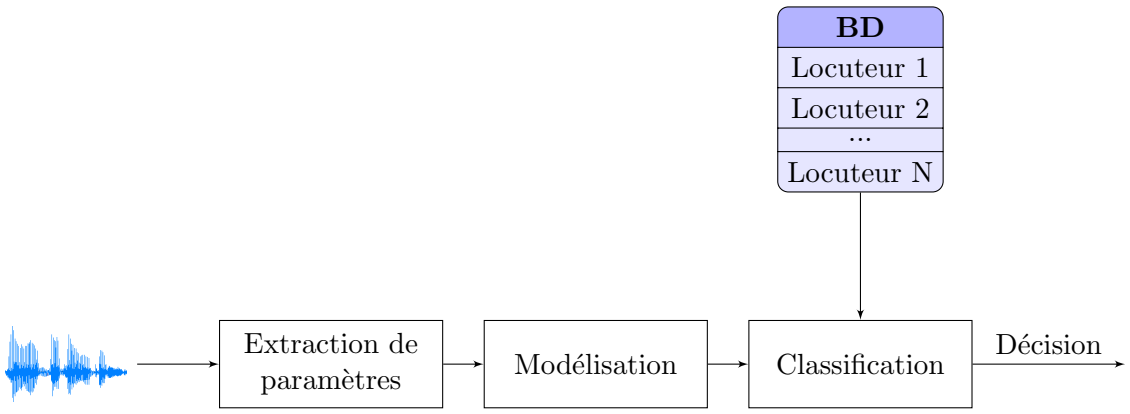



FIGURE 1.1 – Architecture traditionnelle des systèmes de reconnaissance de locuteurs.

Décrivons brièvement les méthodes de l'état de l'art qui s'inscrivent dans chacune des étapes de cette séquence.

#### 1.2.2.1 Extraction de paramètres acoustiques

Un signal de parole contient de l'information propre à la voix du locuteur, de l'information relative au conduit vocal (ce qui est produit), ainsi que des dégradations dues à la production, la transmission et l'acquisition du signal. Reconnaître un locuteur nécessite alors dans un premier temps d'extraire des paramètres discriminant un locuteur d'un autre, indépendamment des sons et des conditions d'enregistrement. Il s'agit alors de minimiser les variabilités intra-locuteur et à maximiser les variabilités inter-locuteurs.

La grande majorité des approches existantes s'appuient sur des paramètres bas niveau extraits sur de courtes fenêtres temporelles. En effet, le signal de parole ne peut être considéré stationnaire que sur de petites fenêtres temporelles, d'environ 25 millisecondes, entre deux articulations. Le vecteur des échantillons dans chaque trame peut alors être représenté par un ensemble de paramètres : les plus populaires en traitement automatique de la parole sont les coefficients cepstraux utilisant l'échelle Mel (MFCC, pour « *Mel Frequency Cesptral Coefficients* ») [DM80]. Ces coefficients représentent une projection dans un espace pseudo-temporel qui décorrèle les informations issues des cordes vocales de celles issues du conduit vocal. Le processus de génération de ces coefficients est traité en détail dans le chapitre 2. Près de 40 ans après leur introduction, ils constituent encore la référence dans l'extraction de paramètres. Parmi les variantes de ces coefficients, les LPCC [RJ93] remplacent l'estimation du spectre du signal par transformation de Fourier par une analyse par prédiction linéaire (LPC pour « *Linear Predictive Coding* »), qui a la propriété d'isoler la production des cordes vocales des résonances du tractus vocal. De l'analyse LPC, les PLP (« *Perceptual Linear Prediction* ») sont extraits dans [Her90], eu égard à une analyse psychoacoustique du signal de parole.
Les paramètres ci-dessus sont les plus courants et le lecteur pourra se référer à [Hat+06] pour une présentation plus exhaustive.

La comparaison entre un vecteur de paramètres acoustiques et un modèle de locuteur peut souffrir des variabilités des conditions d'acquisition et de transmission des signaux de parole. Une étape de normalisation permet d'augmenter la robustesse du système de reconnaissance.

En effet, dans le domaine cepstral, l'effet du canal de communication (production, transmission, acquisition) s'exprime comme un offset, une composante continue dans les valeurs des coefficients. Un procédé de soustraction de la moyenne cepstrale (CMS pour « *Cepstral Mean Substraction* ») est alors proposé dans [Fur81] pour harmoniser les distributions des coefficients cepstraux extraits des différents fichiers de test ainsi que d'apprentissage. Cette normalisation peut aussi être étendue à la variance cepstrale (CMVN pour « *Cepstral Mean and Variance Normalization* ») dans [VL98].

Le « *Feature Wrapping* » proposé dans [PS01] effectue une transformation non linéaire des coefficients dans le but de forcer la distribution des coefficients à adopter un comportement gaussien de moyenne nulle.

Le filtrage RASTA (« *RelAtive SpecTrAL* ») [HM94] utilise un *a priori* sur la distribution log-spectrale de la parole, pour supprimer à l'aide d'un filtre passe-bande des variations potentiellement provoquées par un bruit de fond ou du canal de transmission.

#### 1.2.2.2 Modélisation du locuteur

**Modélisation GMM-UBM** Les systèmes de reconnaissance de locuteurs utilisent le plus généralement une modélisation GMM (« *Gaussian Mixture Models* ») [Rey95], soit une somme pondérées de $M$ lois gaussiennes pour représenter la distribution du vecteur de paramètres acoustiques. Un GMM s'exprime alors par les moments (moyenne $\mu$, matrice de covariance $\Sigma$)

de chaque loi gaussienne, ainsi que le poids associé $\omega$.

$$\text{GMM} = \{\mu_i, \Sigma_i, w_i\}_{i=1}^{M} \tag{1.1}$$

En pratique, l'estimation d'un tel modèle nécessite une très grande quantité de données, rarement disponible en conditions réelles. On substituera alors l'estimation directe des paramètres du GMM en adaptation par Maximum A Posteriori (MAP) d'un modèle du monde (UBM pour « *Universal Background Model* ») [RQD00]. Un UBM est un GMM appris préalablement sur de grandes quantités de données, idéalement proches du contexte applicatif. Il représente la répartition des vecteurs de coefficients du locuteur moyen et ses paramètres sont ensuite adaptés aux données du locuteur cible pour générer leurs propres modèles. Cette modélisation constitue la "baseline" des systèmes de reconnaissance de locuteurs. Nous la formalisons dans le chapitre 2.

Les représentations généralement utilisées sont un vecteur de 39 coefficients pour la paramétrisation et des GMM à 1024 composantes. Nous détaillons les composantes de ce vecteur dans le chapitre 2. La taille de la représentation d'un locuteur $i$ est donc de $1 \times 1024$ pour $\omega_i$, $39 \times 1024$ pour $\mu_i$ et également $39 \times 1024$ pour $\Sigma_i$, en utilisant des matrices de covariance diagonales. Pour alléger cette représentation nous pourrons la réduire en un supervecteur de dimensions $39 \times 1024$, exprimé comme le vecteur des moyennes $\mu_i$, normalisé par $\Sigma_i$ et $\omega_i$.

Comme pour la normalisation des paramètres acoustiques, le post-traitement des supervecteurs créés par le système GMM-UBM a été exploré pour en réduire la dimensionnalité et traiter les variabilités du signal. Un modèle d'analyse conjointe de facteurs (JFA pour « *Joint Factor Analysis* ») est proposé dans [Ken+05], sous l'hypothèse qu'un supervecteur $\mathbf{M}_h(s)$ d'un enregistrement $h$ d'un locuteur $s$ peut être exprimé comme une somme d'un supervecteur $\mathbf{M}(s)$ uniquement dépendant du locuteur et d'un autre supervecteur, uniquement dépendant du canal, $\mathbf{U}\mathbf{x}_h(s)$, où $\mathbf{x}_h(s)$ sont les facteurs du canal et $\mathbf{U}$ une matrice de faible rang qui porte les variabilités inter-canaux de transmission. Les décompositions suivantes sont alors réalisées (équation 1.2) :

$$\mathbf{M}_h(s) = \mathbf{M}(s) + \mathbf{U}\mathbf{x}_h(s) \tag{1.2}$$

$$\mathbf{M}(s) = \mathbf{M} + \mathbf{V}\mathbf{y}(s) + \mathbf{D}\mathbf{z}(s) \tag{1.3}$$

où la matrice $\mathbf{D}$ est diagonale, $\mathbf{V}$ rectangulaire à faible rang et $\mathbf{y}(s)$ et $\mathbf{z}(s)$ sont des vecteurs distribués selon une loi normale standard.

Nous présumons alors que $\mathbf{M}(s)$ suit une loi normale de moyenne $\mathbf{M}$, supervecteur de l'UBM, et de matrice de covariance $\mathbf{V}\mathbf{V}' + \mathbf{D}^2$. $\mathbf{V}$ porte les variabilités entre locuteurs, ses colonnes sont alors les vecteurs propres de la voix, et $\mathbf{y}(s)$ sont les facteurs du locuteur, en dimension réduite. La matrice $\mathbf{D}$ porte une information proche du prior dans l'adaptation MAP du système GMM-UBM et $\mathbf{z}(s)$ sont appelés les facteurs communs.

L'exhaustivité de cette décomposition est remise en question dans [Deh+09] où les variabilités inter-locuteurs et inter-canaux sont projetés dans un unique espace et non deux. Cette espace est nommé espace de variabilité totale. La décomposition suivante est alors réalisée :

$$\mathbf{M}_h(s) = \mathbf{M} + \mathbf{T}\mathbf{w}(s) \tag{1.4}$$

avec **T** la matrice de la variabilité totale, rectangulaire et à faible rang, et $\mathbf{w}(s)$, à dimension réduite est appelé i-vecteur.

L'analyse discriminante linéaire probabiliste (PLDA, pour « *Probabilistic Linear Discriminant Analysis* »), initialement dédiée à la reconnaissance faciale [PE07], constitue l'état de l'art pour la comparaison des i-vecteurs. Cette version probabiliste de l'analyse linéaire discriminante modélise les variabilités du locuteur et celles du canal de transmission dans deux sous-espaces séparés. Elle peut être interprétée comme une JFA dans l'espace des variabilités totales.

#### 1.2.2.3 Classification

La classification d'un segment de parole comme étant prononcé par un des locuteurs de la base de données repose sur des scores assignés à chaque locuteur. En modélisation GMM-UBM classique, ce score pour un locuteur $i$ s'exprime généralement comme un rapport de vraisemblance entre l'hypothèse que le segment de parole $\mathbf{y}$ soit bien prononcé par le locuteur $i$, et l'hypothèse que le segment soit prononcé par un autre locuteur, représenté par le modèle du monde. Ce score est généralement exprimé sur une échelle logarithmique (LLR pour « *Log-Likelihood Ratio* »), ce qui permet une simple décision de reconnaissance selon son signe.

$$LLR_i(\mathbf{y}) = log\left(P(\mathbf{y}|\lambda_i)\right) - log(P(\mathbf{y}|\lambda_{UBM})) \tag{1.5}$$

avec $\lambda_i$ le modèle du locuteur et $\lambda_{UBM}$ le modèle du monde.

Malgré les étapes de compensation des variabilités, au niveau de l'extraction des paramètres et de la modélisation, une partie des effets du canal peut subsister et affecter la répartition des scores de classification. Afin d'atténuer leurs variabilités, les scores $s$ sont centrés réduits selon normalisation suivante :

$$s_{norm} = \frac{s - \mu_{imp}}{\sigma_{imp}} \tag{1.6}$$

où $\mu_{imp}$ et $\sigma_{imp}$ sont les paramètres de la distribution, supposée gaussiennes, des distances entre un locuteur cible et un imposteur (un autre locuteur de la base de donnée).

Nous distinguons deux normes principales : la Z-Norm (pour « *Zero Normalisation* ») qui compense les variabilités inter-locuteurs et la T-Norm (« *Test Normalisation* ») qui se concentre sur les variabilités inter-sessions. L'apprentissage peut être alors réalisé hors ligne pour la Z-Norm mais la T-Norm nécessite le segment de test et l'apprentissage des paramètres imposteurs est alors effectué en ligne.

### 1.2.3 Mise en pratique et cadre d'évaluation

#### 1.2.3.1 Campagnes d'évaluation NIST-SRE

Le NIST [7] (« National Institute of Standards and Technology ») conduit une campagne annuelle depuis 1996 et bi-annuelle depuis 2006 d'évaluation de méthodes de reconnaissance de locuteurs (SRE pour « *Speaker Recognition Evaluation* »). Il fournit un cadre d'évaluation composé de jeux de données distincts par campagnes, pose de nouveaux défis (e.g. SRE'16 a introduit de nouvelles langues dans les segments de test) et permet à la communauté scientifique de se positionner par rapport à un état de l'art relativement à jour.

7. `https://www.nist.gov/`

#### 1.2.3.2 Métriques usuelles

Un système de reconnaissance de locuteurs extrait d'un signal de parole un score de confiance, ainsi qu'une décision binaire acceptation/rejet en fonction du score et d'un seuil $\theta$. À $\theta$ fixe il est alors possible d'évaluer les performances d'un système à travers l'analyse des taux de Faux Rejets (FR) et de Fausses Acceptation (FA) :

$$\text{FR} = \frac{\text{nombre de comparaisons faussement rejetées}}{\text{nombre de comparaisons cible}} \tag{1.7}$$

$$\text{FA} = \frac{\text{nombre de comparaisons faussement acceptées}}{\text{nombre de comparaisons imposteur}} \tag{1.8}$$

**Courbe DET.** Une valeur élevée de $\theta$ imposera une grande exigence pour l'acceptation d'une comparaison test/cible et aura pour conséquence un taux de FR élevé et un taux de FA bas. Inversement, une faible valeur de $\theta$ relaxera le système et provoquera un taux de FR bas et un taux de FA élevé. Les performances du système peuvent alors être représentées exhaustivement dans le plan FA-FR par une courbe, fonction de $\theta$, appelée Detection Error Tradeoff (DET) [Mar+97]. Un exemple de courbe DET est illustré en figure 1.2.

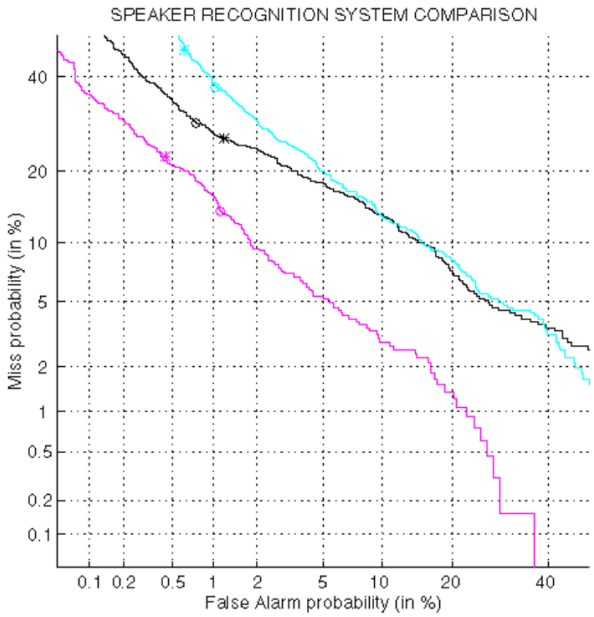



FIGURE 1.2 – Exemple d'une courbe DET : en abscisses le taux de Fausses Acceptations et en ordonnées le taux de Faux Rejets. Figure extraite de [Mar+97].

**Taux Égal d'Erreur (EER pour Equal Error Rate).** Par souci de concision, nous pouvons extraire de cette courbe une valeur caractéristique permettant de comparer plusieurs systèmes à l'aide d'un simple indice de performance. Ce point est situé à l'intersection de la courbe DET et de la courbe d'équation $y = x$, donc à la valeur de $\theta$ pour laquelle FA = FR. Cette mesure est particulièrement pertinente pour des évaluations indépendantes d'une application particulière, un poids identique étant associé aux deux types d'erreur.

#### 1.2.3.3 Outils pour l'implémentation des systèmes de reconnaissance de locuteurs

L'implémentation directe de systèmes de reconnaissance de locuteurs au niveau de l'état de l'art est inenvisageable pour nous du fait de la complexité du système à mettre en place. Nous présentons ici les boîtes à outils les plus couramment utilisées dans la littérature.

**ALIZE.** Il s'agit d'une plate-forme open-source en C++ dont la dernière version [Lar+13] contient les méthodes récentes de l'état de l'art. Largement employée, elle présente une simplicité d'utilisation mais la modification de son architecture est complexe. Par son ancienneté, nous pouvons la considérer comme baseline des outils de reconnaissance de locuteurs.

**SideKit.** Présenté dans [LLM16], cette boîte à outils en Python facilite le prototypage d'un système complet de reconnaissance de locuteurs, à l'inverse d'ALIZE qui nécessite un programme externe pour l'extraction des paramètres et la visualisation des résultats. Elle dispose des méthodes les plus populaires de normalisation (CMS, CMVS, feature wrapping, filtrage RASTA) comme de modélisation et de classification (GMM-UBM, JFA, PLDA, représentation i-vecteurs).

**Kaldi.** Initialement destiné à la reconnaissance de la parole, Kaldi [Pov+11] est une boîte à outils, développée en C++ et adaptée à la recherche qui jouit d'une grande popularité, notamment due au support des réseaux de neurone profonds. Nous retrouvons de nombreuses composantes communes aux tâches de reconnaissance de la parole et de locuteurs, notamment l'emploi des i-vecteurs, motivant ainsi son utilisation.

**MSR.** Il s'agit d'une boîte à outils MATLAB qui permet la mise en place d'un système de reconnaissance de locuteurs basé sur des i-vecteurs couplés à un classifieur PLDA [SSH13].

**SPEAR.** Basée sur le package Python Bob [Anj+12], destiné à l'apprentissage automatique et au traitement du signal, la boîte à outils SPEAR [KESM14] est développée en C++ et Python, proposant un bon compromis performance/accessibilité. Elle est cependant plus axée utilisateurs que développeurs et son back-end est ainsi assez complexe à modifier.

### 1.2.4 Tendances et positionnement

#### 1.2.4.1 Tendances actuelles en reconnaissance du locuteur

Le consortium I4U fait l'inventaire dans [Lee+17] de 17 systèmes de reconnaissance de locuteurs parmi les plus performants lors de la dernière campagne d'évaluation NIST-SRE, en 2016, dont les données d'évaluations et de test comportent pour la première fois plusieurs langues différentes. De cet inventaire il est possible de dégager les tendances actuelles en reconnaissance du locuteur :

- i-Vecteurs et PLDA : l'ensemble des 17 systèmes ont utilisé une approche par i-Vecteurs, la plaçant comme standard actuel en reconnaissance de locuteurs indépendante du texte. Leur dimension varie entre 400 et 600 selon les systèmes. En outre, hormis une approche utilisant une machine à vecteurs de support (SVM), un classifieur PLDA est utilisé dans tous les systèmes pour traiter les variabilités inter-sessions,
- extraction de paramètres : les paramètres MFCC restent toujours très majoritaires, et à dimension croissante, formant un vecteur contenant les coefficients, leurs dérivées et

dérivées secondes le plus souvent de dimension 60 (contre généralement 39 dans les campagnes précédentes),
— le *deep learning* en reconnaissance de locuteurs : incorporé dans 6 des 17 systèmes, il s'intègre à plusieurs niveaux, que ce soit sur le vecteur de paramètres, directement appris sur le signal brut ou couplé à des MFCC, ou en remplacement de la modélisation GMM-UBM [Lei+14].

#### 1.2.4.2 Choix pour nos investigations menées dans le cadre de ces travaux

Les systèmes utilisés dans les campagnes d'évaluation visent à traiter des signaux de parole en conditions de plus en plus difficiles mais nécessitent également des traitements informatiques assez conséquents. Les conditions de notre applications sont bien plus favorables et ne justifient alors pas leur emploi. Par souci de concision, d'implémentation disponible et de simplicité calculatoire et de mise en place, nous privilégions alors une approche classique par GMM-UBM sur un vecteur contenant des MFCC et ses dérivées premières et secondes et de taille totale 39, ainsi que des normalisation CMS pour les descripteurs et ZT-norm pour le scoring.

## 1.3 Signature Visuelle pour la ré-identification de personnes

Nous présentons dans un premier temps la problématique de la ré-identification visuelle, ses enjeux et les principaux défis sous-jacents. Puis, un bref état de l'art est proposé. La section se termine par une description des principales bases publiques d'images et les métriques usitées dans ce contexte.

### 1.3.1 Constats sur la ré-identification

Cette problématique a été initiée dans [GSH06] et largement investiguée depuis. Plusieurs enquêtes synthétisent ces nombreux travaux [Sat13 ; BGS14 ; ZYH16].

**Définition.** Dans [GSH06], la ré-identification est définie comme la capacité à « *déterminer si un individu donné a été observé précédemment au sein d'un réseau de caméras* ». Cette recherche peut également être étendue à une unique caméra, à instants différés. La ré-identification se décompose généralement en deux étapes :
— caractérisation de signature supposée discriminante et invariante aux conditions de prise de vue,
— appariement entre signature de la cible courante observée et signature(s) de la Base de Données (BD) via une mesure de similarité.

Son synoptique, illustré figure 1.3, se décompose comme suit : segmentation/détection de personne(s) dans l'image courante, puis caractérisation du descripteur à partir de région image d'intérêt (RoI) [8]. Le processus d'appariement (donc de ré-identification) s'effectue alors via une mesure de similarité entre ce descripteur ainsi extrait et la base de données des descripteurs relatives aux cibles préalablement perçues.

La détection visuelle de cibles est largement investiguée dans la communauté Vision et sort des objectifs de cette thèse. Nous privilégierons quelques détecteurs existants. ces choix seront motivés dans le chapitre suivant.

8. On parlera de RoI ou d'échantillon.

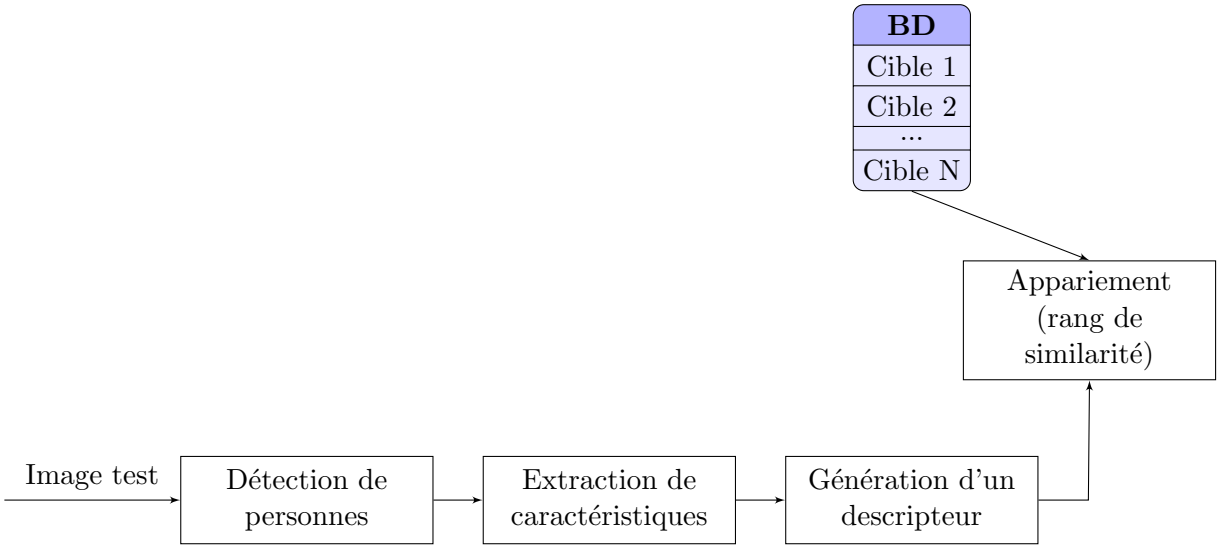



FIGURE 1.3 – Synoptique d'un système de ré-identification traditionnel.

**Variabilité des prises de vue, un verrou clé en ré-identification.** Le processus de ré-identification doit rester robuste aux fortes variabilités de prises de vue. Citons ci-après quelques exemples.

**Variabilité des cibles perçues.** Elle est induite par les situations caméras/cibles, la diversité des postures humaines observables, etc.

**Variations d'illumination.** Elles sont en premier lieu observées pour des environnements non contrôlés. Sur la durée d'une session d'acquisition, ces conditions d'éclairage peuvent varier. Ceci est accentué si la ré-identification est relative à des caméras distinctes, leur rendu est alors lié aux conditions locales d'illumination. Dans ces divers cas, le descripteur doit être robuste à ces variations d'éclairage.

**Occultations partielles.** Les occultations partielles de toute cible sont récurrentes dans les environnements humains : mobilier, architecture de la pièce, autres cibles.

**Variabilité d'apparence.** Le rendu image pour deux caméras distinctes d'un même individu est *a priori* très différent. Des changements de balance des blancs, de contraste ou de saturation peuvent ainsi provoquer de grandes disparités dans la distribution colorimétrique des images, ce dont un descripteur robuste doit pouvoir s'affranchir.

Ces constats motivent notamment la gestion de descripteurs issus de plusieurs échantillons (images). Ainsi, le descripteur agrège les informations issues de plusieurs images successibles de la cibles, extraites de détections appariées dans le flux vidéo.

### 1.3.2 Bref état de l'art en ré-identification

Vous trouverez ci-après un aperçu des méthodes usuelles de ré-identification de la littérature, et notamment un focus sur : (i) les descripteurs, puis (ii) les mesures de similarité entre descrip-

teurs, qui sont, comme évoqué précédemment, des choix clés lors de l'implémentation.

#### 1.3.2.1 Descripteurs

Nous distinguons classiquement les descripteurs basés sur des Points d'Intérêts (PI), soit la recherche de points particuliers, robustes au mouvement, généralement sur le contour de la cible, des descripteurs basés sur des régions d'intérêts (RoI), soit les zones de l'image qui englobent les cibles. Citons ici les travaux de Ghessairi [GSH06] et de Hamdoun [Ham+08] qui exploitent les PI pour la ré-identification. Une méthode par PI est plus robuste aux occultations mais traite une quantité d'information plus faible que les RoI. Les descripteurs associés s'expriment le plus souvent sous forme de distributions de symboles (niveaux).

Citons ici les travaux de [GBT07] dans lesquels le descripteur ELF (« Ensemble of Localized Features ») est une somme pondérée de diverses composantes : histogrammes RGB, YCbCr et HS, filtres de Gabor et de Schmid. La RoI image segmentant une cible/personne est divisée en 5 bandes, représentées chacune par ces 10 indices. Les poids associés à chaque composante du descripteur sont estimés sur un ensemble de données d'apprentissage à l'aide de l'algorithme Adaboost, qui permet d'isoler les indices les plus discriminants. Les canaux H et S (teinte et saturation) obtiennent les poids les plus forts, validant la robustesse de l'espace couleur HSV sur l'espace RVB.

[SD09a] propose un vecteur de caractéristiques similaires, ici à 12 dimensions, qui est extrait des images. Celui-ci est alors réduit en un vecteur de dimension inférieure par analyse PLS (« *Partial Least Squares* »). Cet outil statistique maximise la séparabilité des classes, à l'image de l'Analyse en Composante Principale (ACP), mais conserve l'identité des classes.

Dans [Far+10], la RoI contenant la cible est partitionnée en tête, torse, et jambes, suivant les axes de symétrie et d'antisymétrie de la silhouette d'une personne. Trois descripteurs locaux sont extraits du torse et des jambes de la cible : des histogrammes HSV pondérés par un noyau gaussien centré sur l'axe vertical de symétrie de la silhouette, des régions stables en couleurs (MSCR) [For07] et des patch récurrents de texture (RHCP) [Far+10]. Ce descripteur est nommé SDALF (« *Symetry Driven Accumulation of Local Features* ») et est détaillé dans le chapitre suivant. Les composantes du descripteur sont illustrées sur la figure 1.4.

#### 1.3.2.2 Mesures de similarité entre descripteurs

La littérature propose ici de nombreuses mesures de distances entre vecteurs de descripteurs. Parmi les plus usitées, citons la distance de Bhattacharyya pour comparer des distributions (histogrammes) couleurs. Dans la même veine, citons la distance de Kolmogorov qui, reposant sur la comparaison d'histogrammes cumulés, est certes plus robuste mais reste sensible à l'ordonnancement des couleurs et pour un coût de calcul supérieur.

Une autre distance adaptée aux problèmes de classification est la distance de Mahalanobis, qui effectue une transformation linéaire globale des données d'entrée dans le but d'en éliminer les composantes les plus dispersées. Cette distance requiert le calcul préalable de la matrice de covariance, matrice symétrique semi-définie positive. Dans [WS09], le classifieur LMNN (« *Large Margin Nearest Neighbor* ») apprend une matrice qui minimise les distances entre les images du set d'apprentissage et ses $K$ plus proches voisins portant le même label, et qui les maximise lorsque les labels sont différents. Cette méthode est développée dans [Dik+11] en ajoutant une étape de rejet si les voisins sont trop éloignés. [ZGX11] utilise lui un cadre probabiliste pour maximiser la probabilité qu'une paire d'appariements corrects ait une distance inférieure à celle

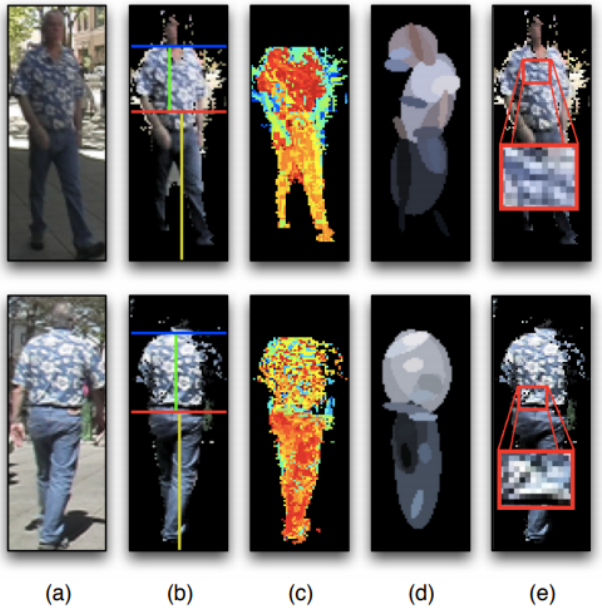



FIGURE 1.4 – Descripteur SDALF : (a) images brutes, (b) partition de la silhouette segmentée, (c) histogrammes HSV, (d) MSCR, et (e) RHCP [Far+10].

d'une paire d'appariements incorrects (PRDC pour « Probabilistic Relative Distance Comparison »).

Parmi les méthodes usuelles, KISSME [Kö+12] apprend une distance de Mahalanobis à partir d'un rapport logarithmique de vraisemblances entre l'hypothèse de dissimilarité entre $x_i$ et $x_j$ et l'hypothèse de similarité, projetée dans l'espace des différences ($x_{ij} = x_i - x_j$) dans lequel les vraisemblances ont un comportement gaussien de moyenne nulle.

Les travaux de [GBT07] ont été étendus dans [Pro+10] où l'intégration de machines à vecteurs de support (SVM) dans la classification transforme le bloc décisionnel basé sur des minimisation de distances en recherche de rang.

Enfin, les avancées récentes de l'apprentissage profond en classification ont initié la définition de métriques spécifiques pour la ré-identification. L'apprentissage d'une métrique DML (« *Deep Metric Learning* ») est proposé dans [Lei+14] à partir des pixels de l'images et qui extrait des descripteurs couleurs et textures puis apprend leur similarité dans un unique « framework », basé sur des réseaux de neurones siamois.

#### 1.3.2.3 Choix pour nos investigations futures

Le choix du (ou des) détecteur(s) visuel(s) sera discuté dans le chapitre suivant. Nous nous appuyons par ailleurs sur des descripteurs et mesures de similarité usuels. Pour les raisons évoquées, nous nous focalisons sur des descripteurs basés RoI. Eu égard à (i) ses performances générales, et (ii) sa robustesse aux variabilités évoquées, nous privilégions le descripteur SDALF. Pour rappel, il repose sur des statistiques relatives à une discrétisation de l'espace des symboles (niveaux) permettant ainsi une signature compacte/compressée, induisant alors des gains CPU lors de leur manipulation. Ce descripteur sera donc détaillé dans le chapitre suivant.

Pour la mesure de similarité entre paires de descripteurs, nous considérons la distance de

Bhattacharrya ; celle-ci, très usitée, est aussi rapide à calculer.

Dans notre contexte, la philosophie « plug and play » des capteurs impose de simplifier au maximum les étapes de déploiement des capteurs. Les capteurs seront étalonnés géométriquement via une étape hors ligne et sommaire d'étalonnage. Cette étape permettra alors d'inférer les mouvements des cibles dans le plan du sol donc tirer parti de leurs mouvements réels, et non leurs mouvements apparents (souvent plus ambigus...) dans les plans capteurs. Nous détaillons ci-après quelques métriques et bases publiques pour évaluer les fonctionnalités visuelles seules en vue d'une démarche incrémentale de validation.

### 1.3.3 Evaluations : métriques et bases de données en vision

#### 1.3.3.1 Métriques usuelles

**Courbes CMC.** La ré-identification de personnes peut être traitée comme un problème de rang. Sous l'hypothèse d'un environnement fermé, l'ensemble des images test correspond à un sous-ensemble de la base de données des cibles. Ainsi pour chaque image, nous pourrons classer les cibles de la base de données par distance croissante image test/cible.

Les méthodes de ré-identification sont alors couramment évaluées par des courbes CMC, pour « *Cumulative Matching Curves* » qui représentent le taux de ré-identification en fonction des son rang $r$, i.e. la probabilité de retrouver la bonne cible à apparier à une image test parmi les $r$ cibles les plus probables. L'aire normalisée sous une courbe CMC (nAUC pour « *normalized Area Under Curve* ») peut également être utilisée comme indice de performance de la ré-identification en une seule dimension. Enfin, certaines valeurs particulières peuvent être extraites des courbes CMC, notamment la valeur au rang 1. Un exemple de courbe CMC est illustré en figure 1.5.

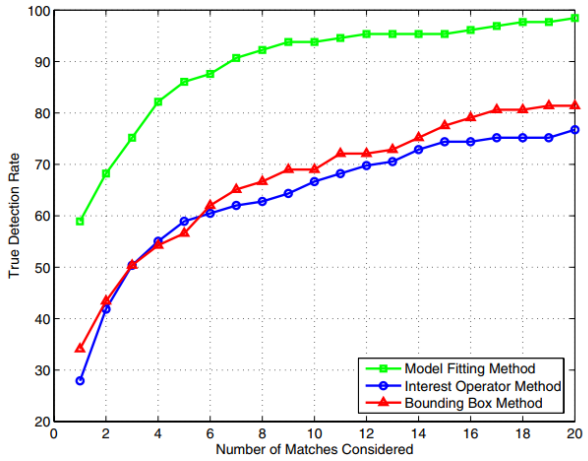



FIGURE 1.5 – Exemple d'une courbe CMC : taux de ré-identification *vs.* rang $r$ [GSH06].

#### 1.3.3.2 Bases de données

Les nombreux travaux sur la ré-identification de personnes ont conduit à la création de bases publiques pour « benchmarks ». Citons ici trois exemples : VIPeR, ETHZ, et i-LIDS.

**VIPeR (Viewpoint Invariant Pedestrian Recognition).** Cette base de données est inhérente à [GBT07] pour de la ré-identification au sein d'une paire de caméras, en mono-échantillon :

seules deux occurrences de chaque cible y sont présentes. Elle contient 632 paires d'images de piétons, capturées avec des points de vue différents et soumis à des variations de luminosité et à des occultations sévères, ce qui constitue alors une base de données complexe (voir figure 1.6).

FIGURE 1.6 – Échantillons issus de la base de données VIPeR [GBT07].

**ETHZ-REID (Eidgenössische Technische Hochschule Zürich).** Présentée dans [ELG07], cette base de données contient des images extraites de caméras mobiles, changeant ainsi les points de vues relatifs et suivant des piétons en environnement dense. À l'inverse de la base VIPeR, elle est multi-échantillons (voir figure 1.7). Elle est ensuite divisée dans [SD09b] en trois sous-ensembles : ETHZ1 (83 personnes, 4857 images), ETHZ2 (35 personnes, 1936 images) et ETHZ3 (28 personnes, 1762 images). Les images d'une même cible sont ici toutes extraites de la même caméra, pour les données dédiées à l'apprentissage comme celles dédiées aux tests, la rendant beaucoup moins contraignante que VIPeR. Ainsi, elle montre ses limites et des méthodes récentes [Lis+15 ; Mar+15] approchent les 100% de taux de ré-identification au rang $r = 1$.

FIGURE 1.7 – Échantillons issus de la base de données ETHZ-REID [ELG07].

**i-LIDS (Imagery Library for Intelligent Detection Systems).** Située à l'intersection de VIPeR et d'ETHZ, les données de cette base privée [Pro+10] sont multi-échantillons et issues de plusieurs caméras ambiantes, dans un terminal d'aéroport aux heures de pointes. La taille de la base est de 476 images pour 119 cibles. Les caméras sont à champs disjoints (NOFOV pour « *Non Overlapping Field of View* ») et les images sont soumises à de fortes variabilités d'illumination et à un grand nombre d'occultations (voir figure 1.8).

FIGURE 1.8 – Échantillons issus de la base de données i-LIDS [Pro+10].

**Synthèse.** La base de données i-LIDS répond *a priori* à nos attentes, mais elle est privée et nous n'avons pas pu y avoir accès. La base VIPeR est mono-échantillon et ainsi inadaptée à notre problème. Nous choisissons alors d'utiliser les données ETHZ pour valider nos investigations côté vision. Elle correspond certes partiellement avec notre contexte (caméra mobile en extérieur), mais les images extraites présentent des similarités satisfaisantes avec notre scénario expérimental.

## Conclusion

Ce chapitre a détaillé en préambule la problématique, ses tenants et aboutissants. Les méthodes existantes ont été ensuite brièvement énumérées : (i) reconnaissance de locuteurs, tâche utile à la génération de la signature audio d'une cible, et (ii) en ré-identification de personnes, pour la génération de la signature vidéo.

En reconnaissance de locuteurs, en 1.2, nous avons présenté la chaîne de traitement sur laquelle se greffent les nouvelles méthodes de la littérature : extraction de caractéristiques, modélisation, et classification. Le défi majeur est la séparabilité des effets du locuteur, du texte, et du canal de transmission dans un signal de parole. La paramétrisation est quasi systématiquement réalisée par des MFCC mais plusieurs normalisations de ces paramètres sont proposées afin de minimiser les variabilités inter-locuteurs et inter-sessions. La démarche est similaire pour la modélisation, souvent basée sur un système GMM-UBM qui peut ensuite être converti en supervecteur et être projeté dans des sous espaces de dimensions réduites (i-vecteurs), ce qui minimise les variabilités qui affectent la reconnaissance.

Par analogie, nous avons listé les verrous en ré-identification visuelle de personnes puis présenté quelques approches usuelles 1.3, en focalisant sur les descripteurs et mesures de similarité (règle de décision) associées.

Pour chacune des modalités, nous avons également présenté les outils et les bases de données pour des évaluations quantitatives (métriques, boîtes à outils ou campagne d'évaluation) les plus usités dans leurs communautés respectives.

Cet état de l'art, certes non exhaustif, a permis de confirmer/affiner nos choix. Ainsi, notre signature audio privilégie une approche classique GMM-UBM sur des MFCC pour sa simplicité d'intégration. Dans des contextes contraignants, elle est dépassée par les meilleures approches (couplage i-vecteurs et PLDA) mais convient parfaitement à notre contexte applicatif. Côté

vision, notre signature privilégie le descripteur SDALF car il est largement usité de par sa robustesse aux variabilités diverses de prises de vue. Ce descripteur statistique a trois composantes : deux utilisent la distance de Bhattacharryya et la troisième une mesure personnalisée. Enfin, nous utiliserons la base de données publiques ETHZ pour valider notre signature vidéo.

En s'appuyant sur les choix listés et justifiés ici, nous expliciterons les signatures audio et vidéo et (surtout) leur association dans notre contexte applicatif, dans les chapitres suivants.

# Chapitre 2

# Signature audio, signature vidéo : concepts et techniques

**Sommaire**



## Préambule

Dans l'introduction générale, nous avons relevé le côté novateur de notre problématique, dans son aspect de fusion audiovisuelle en contexte ambiant. Il est ainsi difficile de trouver des bases de données publiques correspondant précisément à notre contexte applicatif.

Parmi les bases de données audiovisuelles existantes, certaines sont dédiées à l'identification du locuteur, avec de prise de vue face caméra, comme VidTIMIT [SL09] et sortent de notre contexte. Parmi les bases de données en contexte plus ambiant, nous pouvons relever la base AV16.3 [LOGP05] qui instrumente une salle de réunion avec 3 caméras et deux chaînes de 8 microphones. Si ces acquisitions se placent en contexte ambiant, l'aspect épars de l'instrumentation n'est cependant pas retrouvé ici. De plus la salle est plus petite que dans notre contexte (salle pédagogique) et limite ainsi le nombre d'usagers ainsi que leurs interactions.

### Création d'une base de données audiovisuelles personnelle

Face à l'absence de données publiques correspondant à notre problématique, nous avons choisi de réaliser des acquisitions audio-vidéo, en environnement contrôlé dans un premier temps, afin d'établir aisément une vérité-terrain en terme de locuteurs et de localisation. Nous avons équipé une salle de l'Université de caméras et microphones, dans la configuration représentée en vue zénithale sur la figure 2.1 (a).

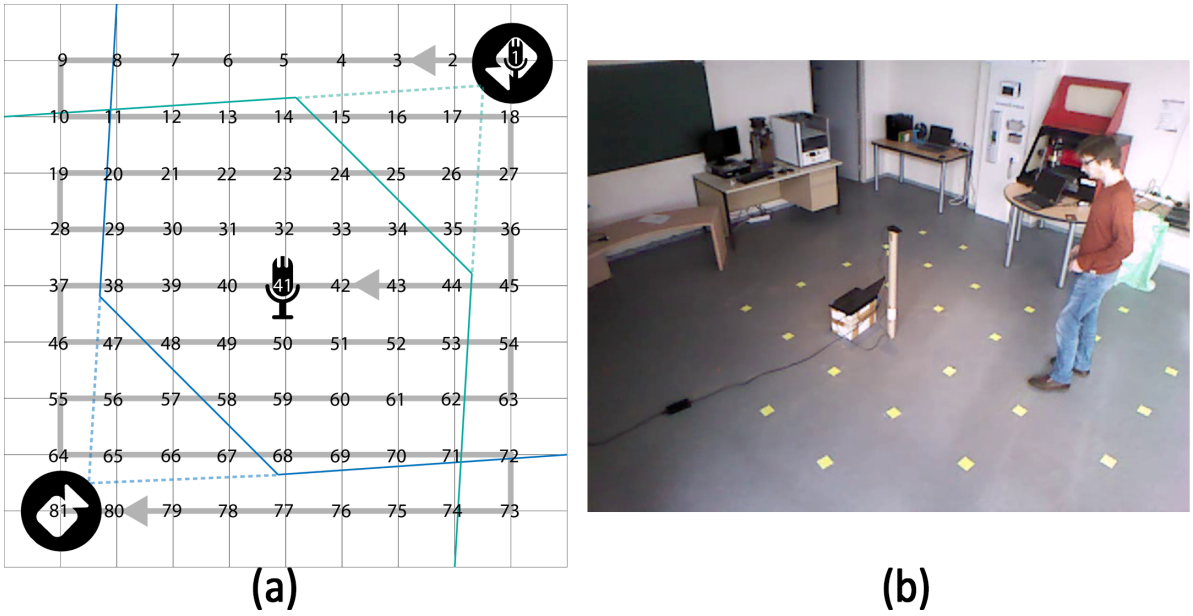



FIGURE 2.1 – Configuration de notre plate-forme expérimentale (a), image extraite de la caméra 1 (b).

**Instrumentation** Nous avons placé des marquages, régulièrement espacés de 62 cm sur le sol de la salle d'expérimentation, de dimensions approximatives 6 m par 6 m. 81 positions sont alors extraites, formant une discrétisation spatiale de l'espace d'acquisition. Deux microphones sont installés, respectivement dans un coin de la salle et au centre de la pièce, ainsi que deux caméras dans deux coins opposés de la salle, à plus de 2 mètres de hauteur. Dans notre contexte de capteurs épars, tous les capteurs n'ont pas vocation à être utilisés simultanément, cette installation nous permet alors d'explorer plusieurs configurations. Cette installation est représentée en figure 2.1 (a).

Les capteurs utilisés sont les suivants :

— microphones MXL AC-404, à faible coût (inférieur à 100€) et avec interface en USB, dédiés à des applications type visioconférence. Enregistrement monophonique à 16 kHz,

— caméras Microsoft Kinect, dont nous utilisons uniquement les canaux RGB. Résolution de $640 \times 480$ pixels et enregistrement à 24 images par seconde. Un exemple d'image issue d'une caméra est illustrée en figure 2.1 (b).

**Scenarii d'acquisition** L'objectif de ces premières acquisitions est d'obtenir un jeu de données contrôlées, afin de valider les méthodes développées dans ces travaux dans des conditions simples dans un premier temps. Ainsi, 3 participants, dont les apparences vestimentaires sont distinctes, parcourent successivement le trajet représenté en figure 2.1 (a), à l'aide des marquages au sol.

Nous obtenons ainsi un parcours exhaustif de l'ensemble de la pièce ainsi qu'une vérité-terrain discrète et sans équipement de la localisation des locuteurs.

À chacune des 81 positions de la pièce, les participants diffusent, à l'aide d'une enceinte Bluetooth, un fichier sonore contenant 20 secondes de parole et issu du corpus de parole propre BREF [Lam+]. Chaque participant diffuse un segment de parole différent et prononcé par un locuteur différent, créant ainsi un corpus de parole à plusieurs distances du microphones.

Enfin, afin de simuler des conditions de foule, ces expériences sont reproduites en diffusant un « babble » à l'aide d'une enceinte dans la pièce. Ce bruit est créé par la superposition de 7 segments de paroles, prononcés par des locuteurs différents, décalés temporellement et diffusés en boucle. Nous avons arbitrairement fait varier sa puissance sonore et mesuré ensuite son rapport signal sur bruit (SNR pour « *Signal to Noise Ratio* ») en comparant des trames de bruit seul et des trames mélangeant parole et bruit. Nous obtenons alors les valeurs suivantes : 13.3 dB pour le bruit le plus faible, 6 dB pour le bruit intermédiaire, et -3.4 dB pour le bruit le plus fort.

La base de données ainsi créée sera alors utilisée dans ce manuscrit pour supporter nos diverses évaluations audiovisuelles.

**Métriques d'évaluation** Nous avons présenté dans l'introduction générale un certain nombre de métriques propres aux tâches de ré-identification visuelle de personne et de reconnaissance de locuteurs. Nous rappelons ici quelques métriques usuelles en classification.

Considérons les notations suivantes : $c$ l'événement cible, $FN$ un Faux Négatif, $FP$ un Faux Positif, $TP$ un Vrai Positif, $TN$ un Vrai Négatif, $T(c)$ le nombre de trames où $c$ est présent, $T(nc)$ le nombre de trame où $c$ n'est pas présent. On définit alors les métriques suivantes :

— Taux d'erreur (Error Rate) :

$$ER = \frac{\sum FN + \sum FP}{\sum T(c) + \sum T(nc)}$$

— Rappel : la proportion des solutions pertinentes qui sont trouvées

$$rec = \frac{\sum TP}{\sum TP + \sum FN} = \frac{\sum TP}{\sum T(c)}$$

— Précision : la proportion de solutions trouvées qui sont pertinentes

$$prec = \frac{\sum TP}{\sum TP + \sum FP}$$

— F-mesure : la moyenne harmonique de la précision et du rappel, regroupant en une unique mesure la précision et le rappel

$$F = \frac{2 * rec * prec}{rec + prec}$$

# Introduction

Pour rappel, la ré-identification consiste ici à apparier des observations issues de capteurs audios et optiques. Elle peut s'opérer au sein d'un réseau de capteurs : nous rechercherons, par exemple, à reconnaître dans le champ de vue d'une caméra un individu préalablement perçu par une autre caméra du réseau. La ré-identification peut également intervenir au niveau d'un simple capteur mais à des instants différents : nous rechercherons, par exemple, à regrouper les différentes occurrences d'un locuteur enregistré par un microphone.

Dans notre contexte applicatif, décrit dans le chapitre d'introduction, une salle est équipée de microphones et de caméras ambiants, fixes, dont les positions dans le plan du sol sont connues et à champs partiellement joints, formant ainsi un réseau épars. Afin de respecter une philosophie « plug and play », avec la désactivation possible de certains éléments du réseau, chaque capteur doit pouvoir contribuer individuellement, sans s'appuyer sur un quelconque a priori sur la topologie du réseau de capteurs. Notre objectif est donc d'associer les percepts produits à différents instants et/ou issus de différents capteurs, potentiellement hétérogènes, relatifs à la personne cible transitant au sein du réseau.

Ré-identifier un individu, aussi bien visuellement qu'auditivement, nécessite l'extraction préalable de signatures qui lui sont propres. Sa robustesse est évaluée par sa capacité à reconnaître une cible observée auparavant et par sa capacité à la distinguer d'une autre, elle doit ainsi :

— minimiser les variabilités intra-cibles, soit maximiser $P(obs_i|sign_i)$,
— maximiser les variabilités inter-cibles, soit minimiser $P(obs_i|sign_j)$.

Avec $i$ l'index de la cible observée, $obs_i$, $sign_i$ et $sign_j$ les signatures respectivement observées et apprises des cibles $i$ et $j$ pour $(i,j) \in [1,N]^2, i \neq j$, et $N$ le nombre total de cibles.

## Problématique et verrous scientifiques

La question suivante se pose alors : quelles signatures discriminantes et compactes pouvons-nous privilégier eu égard au contexte de notre étude ? Les caractéristiques audiovisuelles d'un individu inconnu les plus discriminantes a priori sont le timbre de sa voix et certains traits de son visage. Si les flux audio des microphones ambiants en environnement non contrôlé peuvent contenir l'information suffisante à caractériser le timbre du locuteur, un protocole expérimental supervisé est nécessaire à l'extraction de descripteurs visuels du visage et sort donc de notre contexte applicatif. Les scènes observées étant contraintes dans le temps, de l'échelle de la demi-heure à 2 ou 3 heures, nous pourrons alors nous appuyer sur l'hypothèse que les participants conservent les mêmes vêtements pour construire une signature visuelle basée sur l'apparence du corps complet.

## Synoptique de notre approche

Nous présentons ici les étapes menant à la conception d'un système d'apprentissage d'une signature audiovisuelle d'un individu à partir de flux audio et vidéo extraits respectivement d'un microphone et d'une caméra. Les caractéristiques que nous modélisons ne présentent cependant pas de corrélation. L'apparence et le timbre de voix d'un individu sont en effet indépendants, ainsi les signatures audio et vidéo seront produites séparément, puis associées par fusion tardive lorsque les observations d'où elles sont extraites seront localisées au même emplacement ou suffisamment proches. L'architecture complète du système est présentée sur le schéma-bloc 2.2. Sa composante bleue illustre la génération des modèles audio et vidéo et fera l'objet de ce

chapitre, à travers les sections 2.1 et 2.2 respectivement. La localisation multimodale de la personne détectée (en vert sur le synoptique), sera traitée dans le chapitre suivant, en section 3.1, préalablement à la fusion des signatures (partie rouge), traitée en section 3.2.

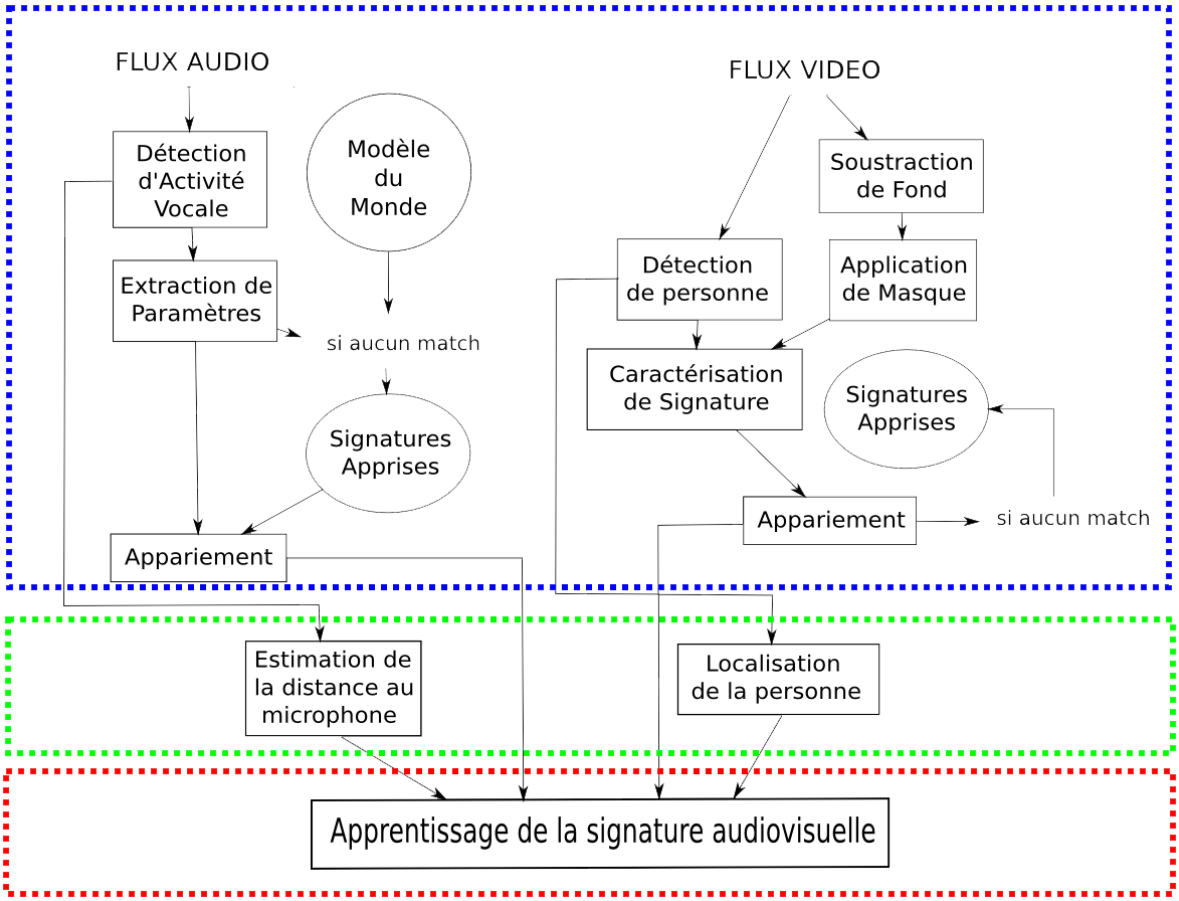



FIGURE 2.2 – Synoptique de notre système d'apprentissage d'une signature audiovisuelle de personne.

## 2.1 Signature Audio

Cette section est consacrée à la présentation des concepts et outils qui composent la chaîne de traitement sonore, illustrée sur le diagramme 2.3, utilisée pour générer la signature audio d'une personne. Elle suit une approche séquentielle, et les blocs la composant sont explicités dans les sous-sections ci-après. Une première étape, la détection d'activité vocale, segmente le flux audio capté par le microphone pour ne conserver que des segments de parole. Elle sera traitée en section 2.1.1. De ces segments sont ensuite extraits des descripteurs, explicités en section 2.1.2, portant l'identité du locuteur. Enfin, un modèle est appris à partir de ceux-ci, formant la signature audio du locuteur. L'apprentissage du modèle sera détaillé en section 2.1.3.

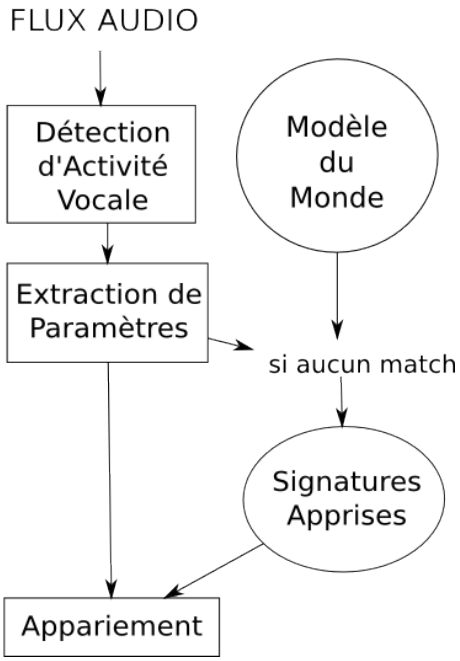



FIGURE 2.3 – Chaîne de traitement pour la génération d'une signature audio.

### 2.1.1 Détection d'Activité Vocale

Nous entendons par détection d'activité vocale (VAD, pour « Voice Activity Detection ») l'extraction des zones temporelles d'un signal audio comportant de la parole ou du chant. Par abus de langage et dû à son grand nombre d'applications, en amont des systèmes de codage, de synthèse, d'analyse ou de reconnaissance automatique de la parole, elle est généralement utilisée pour définir la détection des zones de parole. Les résultats attendus s'expriment via l'équation 2.1 et sont illustrés sur la figure 2.4 :

$$\mathrm{VAD}(i) = \begin{cases} 1 & \text{si la trame } i \text{ contient de la parole} \\ 0 & \text{sinon} \end{cases} \qquad i = 1, \ldots, T \tag{2.1}$$

avec $T$ le nombre total de trames extraites du signal audio original.

De nombreuses techniques ont été proposées, qu'il est possible de répartir en deux catégories : les approches basées apprentissage automatique, et les approches basées traitement de signal.

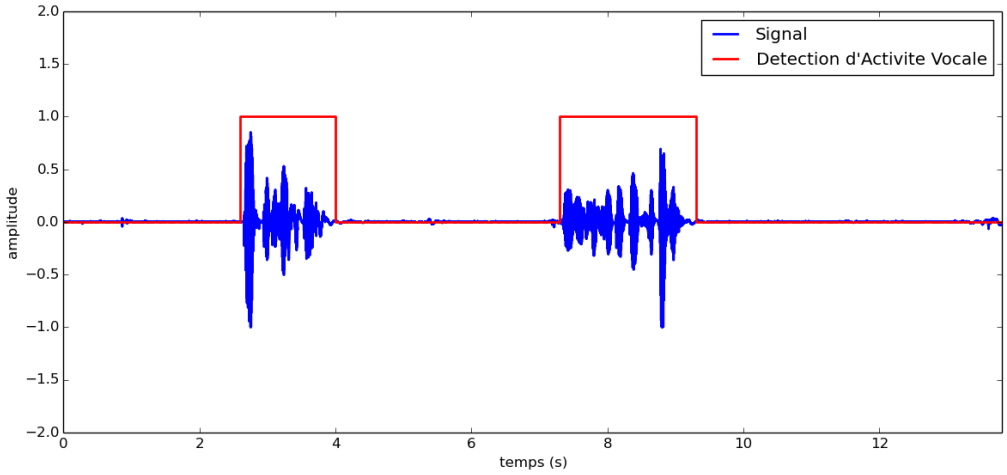



FIGURE 2.4 – Exemple de sortie d'un détecteur d'activité vocale sur un fichier audio de 14 secondes, contenant deux segments de parole. Le signal, échantillonné à 16 kHz, a été analysé par en utilisant des trames de 16 ms.

#### 2.1.1.1 Approches fondées sur l'apprentissage automatique

Ces approches nécessitent en amont une description du signal de parole le plus souvent réalisée par l'extraction de paramètres MFCC (voir 2.1.2).

Parmi les méthodes de classification les plus classiques, les machines à vecteurs de support (SVM), ainsi que ses extensions, comme les SVM à noyaux multiples, ont été appliquées à la détection d'activité vocale [WZ11]. Une alternative [SKS99] consiste à entraîner des modèles statistiques de trames contenant de la parole et d'autres contenant uniquement du silence ou du bruit, sous forme de mélanges de lois gaussiennes (GMM). En phase de test, les nouvelles trames sont classées en fonction du rapport de log-vraisemblance à ces deux modèles. Enfin, la récente popularité de l'apprentissage profond a amené de nouvelle méthodes basées sur des réseaux de neurones récurrents (RNN) et convolutifs (CNN). Une évaluation de leur robustesse à différents niveaux de bruits est réalisée en [TGY16].

Ces méthodes étant mal adaptées au contexte de notre étude par leur complexité et surtout la nécessité d'un corpus d'apprentissage, nous ne nous sommes ici focalisés que sur quelques travaux représentatifs dans ce domaine.

#### 2.1.1.2 Approches fondées sur le traitement de signal

Bien que leurs performances en conditions bruitées puissent être surpassées, les méthodes basées sur l'extraction d'indices acoustiques bas ou haut niveau offrent une alternative élégante et tout à fait adaptée à de nombreuses applications. Pour l'extraction de ces paramètres, le signal original est découpé en trames de généralement 16 ms, avec possible recouvrement entre elles. Les notations suivantes seront alors employées :

— $i$ l'index de la trame considérée,
— $N$ le nombre total d'échantillons de signal dans la trame,
— $x_n(i)$ le $n$-ième échantillon de la trame $i$ du signal $x$.

Nous présenterons ci-après les mesures les plus populaires et les évaluerons ultérieurement sur la base de données constituée dans le préambule de ce chapitre.

**Énergie à court terme du signal** Un des indices les plus instinctifs est l'énergie à court terme du signal, calculée sur chaque trame du signal. Elle s'exprime comme suit :

$$E(i) = \sum_{n=1}^{N} x_n^2(i) \tag{2.2}$$

Si elle est peu robuste à des distinctions parole/musique ou parole/bruits environnementaux, elle peut toutefois se révéler performante dans des scenarii ne contenant a priori aucun autre signal aussi énergique, et de nombreux systèmes l'utilisent encore pour sa simplicité.

**Le ZCR** Le taux de passage par zéro (ZCR pour « Zero Crossing Rate ») est un autre indice temporel qui observe les alternances d'amplitude autour de la valeur centrale du signal, en général zéro. En raison de sa nature aléatoire, une trame de bruit possédera un ZCR plus élevé qu'une trame de signal voisé. Il est calculé de la manière suivante :

$$ZCR(i) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}\{x_{n-1}(i)x_n(i) < 0\} \tag{2.3}$$

avec $\mathbb{1}$ la fonction indicatrice.

L'énergie et le ZCR comme mesures de détection d'activité vocale ont été exploités de longue date, et les variations de ces descripteurs sur des fenêtre temporelles adjacentes ont également prouvé leur pouvoir discriminant [LJZ01].

**Modulation de l'énergie à 4 Hertz** Ce descripteur, plus haut niveau, exploite des propriétés propres au signal de parole. Le débit syllabique (comprendre le nombre de syllabes prononcées par seconde) s'élève aux alentours de 4 Hz. En considérant qu'une syllabe est généralement composée d'une partie voisée (haute en énergie) et d'une non voisée (faible en énergie), il est possible d'observer cette alternance dans son spectre de modulation [HS85].

Dans le domaine fréquentiel une banque de filtre suivant l'échelle perceptive Mel est appliquée à chaque trame, générant une valeur d'énergie par canal. Les signaux d'énergie par canal ainsi produits sont filtrés par un filtre passe-bande centré en 4 Hz, puis sommés et normalisés par le maximum de l'énergie. Le calcul de la variance sur des fenêtres d'une seconde donne la valeur de la modulation de l'énergie à 4 Hz.

**Autres indices pour la détection d'activité vocale** De même que pour la modulation de l'énergie à 4 Hertz, l'analyse de la modulation de l'entropie permet d'extraire les alternances des segments voisées, à faible entropie, et des segments non voisés, à haute entropie [PSAo02]. Dans le domaine fréquentiel, l'estimation de la fréquence fondamentale, de l'harmonicité et de la divergence spectrale peuvent être des indicateurs efficaces de la présence d'activité vocale. Enfin, l'étude des statistiques d'ordre supérieur dans les résidus des LPC est utilisée dans [NGM01].

#### 2.1.1.3 Évaluations des descripteurs

Les méthodes utilisant la modulation de l'énergie à 4 Hz et la modulation de l'entropie ont montré de très bonnes performances [Gal+05] par l'équipe SAMoVA de de l'IRIT lors de la campagne d'évaluation des Systèmes de Transcription Enrichie d'Émissions Radiophoniques [9] (ESTER). Nous évaluons la performance des méthodes sur une partie du corpus décrit en préambule

9. http://www.afcp-parole.org/camp_eval_systemes_transcription/

de ce chapitre, comportant des alternances de parole à plusieurs distances du microphone et de silence. Les métriques suivantes, détaillées en préambule, seront utilisées : taux d'erreur ($ER$) et F-mesure ($F$). Les résultats sont présentés sur le tableau 2.1 :

TABLE 2.1 – Évaluations des méthodes de VAD sur le corpus.

| | **Métriques** | |
|---|---|---|
| **Méthodes** | $ER$ | $F$ |
| Énergie | 0.37 | 0.73 |
| ZCR | 0.43 | 0.73 |
| Modul. énergie 4 Hz | **0.33** | **0.78** |
| Modul. entropie | 0.34 | 0.77 |

Les meilleures performances sont obtenues avec la détection par la modulation de l'énergie à 4 Hz et nous utiliserons alors cette méthode par la suite. La modulation de l'entropie présente des performances semblables alors que les paramètres plus bas niveau (énergie, et ZCR) se révèlent légèrement moins efficaces. Notons l'importance de l'établissement de la vérité terrain dans le processus d'évaluation : une annotation manuelle aura tendance à regrouper de longs segments de parole (plusieurs secondes) alors que les paramètres ci-dessus sont extraits sur des trames de 16 ms, d'où certaines ambiguïtés d'étiquetage.

Le signal de parole ainsi filtré, seuls les segments contenant de la parole seront traités. Nous nous intéressons maintenant à l'extraction des descripteurs portant l'information du timbre du locuteur dans ces segments.

### 2.1.2 Paramètres pour la reconnaissance du locuteur

La question centrale de la tâche de reconnaissance de locuteur est d'isoler ce qui porte l'identité du locuteur dans le signal de parole. Pouvoir séparer les informations relatives au texte et celles propres au locuteur est nécessaire aussi bien aux systèmes de reconnaissance de locuteurs qu'aux systèmes de reconnaissance automatique de la parole (transcription). Il est donc nécessaire d'en extraire des descripteurs indépendants du texte et robustes aux variations de l'appareil de production de la parole, puis d'en générer des modèles.

#### Calcul des coefficients cepstraux

Du à son caractère articulé, le signal de parole change continuellement et nécessite d'être découpé en segments de 15-30 ms sur lesquels le signal pourra être considéré comme stationnaire et nous pourrons alors extraire des descripteurs bas niveau sur ces trames. S'il reste un domaine de recherche actif, la création de descripteurs bas niveau est dominée par l'utilisation des coefficients cepstraux suivant l'échelle Mel (MFCC, pour Mel-Frequency Cepstral Coefficients) introduits par [DM80]. Le calcul de ces coefficients est détaillé dans le paragraphe suivant. Les coefficients cepstraux peuvent également être extraits sur une échelle linéaire, et non Mel, il s'agit alors de LFCC (« Linear Filter Cepstral Coefficients »).

**MFCC** La conception de ces paramètres repose sur la théorie source-filtre [Fan60], selon laquelle la production de la parole résulte d'un produit de convolution entre une source $g(.)$,

ensemble de générateurs de sons, et un filtre $h(.)$, ensemble de résonateurs qui affectent le son source :

$$x(t) = g(t) * h(t) \tag{2.4}$$

La source est composée d'ondes glottiques, générées par la vibration des cordes vocales et fortement harmoniques, ainsi que de constrictions dans le conduit vocal, qui sont à l'origine de bruit, apériodique. Le filtrage est réalisé par l'ensemble du tractus vocal, composé de plusieurs cavités supra-glottiques, qui amplifie certaines harmoniques du signal source, et est responsable du timbre du locuteur. L'idée derrière l'extraction des coefficients cepstraux d'un signal est d'effectuer une suite de transformations visant à séparer la source du filtre, de faire de la convolution une simple addition.

Le processus de génération des MFCC est présenté sur le diagramme 2.5.

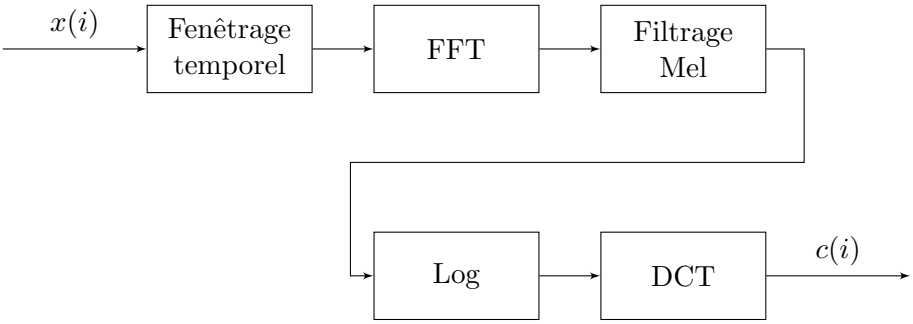



FIGURE 2.5 – Processus de génération des MFCC, notés $c(i)$, depuis une trame de signal $x(i)$.

Les étapes sont les suivantes :

**Fenêtrage et passage dans le domaine de Fourier** Le passage du domaine temporel au domaine fréquentiel par transformée de Fourier rapide (FFT pour « Fast Fourier Transform ») transforme la convolution en produit. L'équation 2.4 devient alors :

$$X(\omega) = G(\omega)H(\omega) \tag{2.5}$$

Un fenêtrage (type Hamming ou Hanning) en amont de la FFT est nécessaire afin de lisser les transitions au début et à la fin de la trame et ainsi limiter l'amplitude des lobes secondaires dans l'estimation du spectre.

**Banque de filtres Mel et passage à l'échelle logarithmique** Le mel est une unité de hauteur de son qui respecte les distances perceptives. L'audition humaine possède une sensibilité aux hautes fréquences plus faibles, notamment à partir de 1000 Hz : à plusieurs centaines de Hertz différentes pourra correspondre une unique hauteur de son perceptible, alors représentée par un mel. La transformation Hertz-Mel se calcule de la façon suivante :

$$m = 1127.\ln\left(1 + \frac{f}{700}\right) \tag{2.6}$$

Cette représentation est exploitée à travers l'application d'une banque de $N$ filtres triangulaires dont les fréquences centrales suivent l'échelle Mel, comme illustré en figure 2.6. Nous

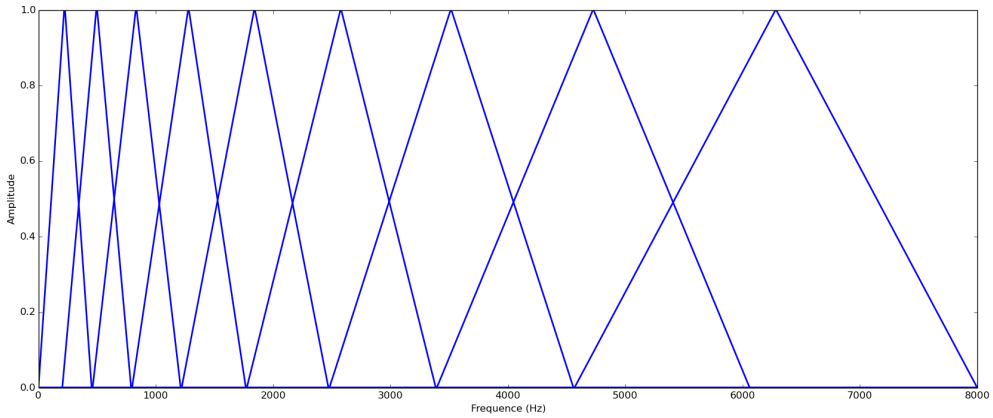



FIGURE 2.6 – Réponse fréquentielle d'une banque de 10 filtres suivant l'échelle perceptive Mel sur l'espace fréquentiel [0-8000Hz].

retenons ensuite une valeur en sortie de chaque canal, somme des composantes énergétiques dans la bande de fréquences correspondante.

En outre, la psychoacoustique a montré la pertinence de l'utilisation d'échelle logarithmique comme échelle de sonie [RD56]. L'humain perçoit mieux les variations de stimuli de faible intensité que ceux de haute intensité. L'échelle linéaire est ainsi inadéquate à la représentation de la perception auditive.

**Transformée en cosinus discrète** La transformée en cosinus discrète (DCT pour « Discrete Cosine Transform ») effectue une reprojection dans un domaine pseudo-temporel. Par sa proximité avec la transformée de Karhunen-Loeve, elle tend à décorréler les log-énergies (en échelle Mel) et génère $K$ coefficients réels :

$$c_k(i) = \frac{1}{K} \sum_{n=1}^{N} \log(MF(n)) \cos\left(\frac{\pi}{N}\left(n - \frac{1}{2}\right)k\right) \tag{2.7}$$

$c_k(i)$ est le $k$-ème coefficient cepstral de la trame $i$, $MF(n)$ l'énergie dans la bande $n$ de la banque de filtres Mel. Les premiers coefficients portent l'information de timbre. Les variations de ces coefficients peuvent également porter de l'information, dû au caractère changeant de la parole. Notre vecteur de paramètres sera classiquement composé des coefficients, de leurs dérivées $\Delta$ et de leurs dérivées secondes $\Delta\Delta$, ainsi que de l'énergie :

— 12 coefficients cepstraux,
— 12 $\Delta$ coefficients cepstraux (vitesse),
— 12 $\Delta\Delta$ coefficients cesptraux (accélération),
— 1 coefficient d'énergie,
— 1 $\Delta$ coefficient d'énergie,
— 1 $\Delta\Delta$ coefficient d'énergie.

Soit un vecteur de taille 39 par trame. Ces coefficients extraits vont ensuite alimenter un modèle statistique qui représentera notre signature audio.

### 2.1.3 Modélisation

Comme présenté en chapitre 1, les modèles de référence en reconnaissance automatique du locuteur sont basées sur des mélanges de lois gaussiennes (GMM) adaptées depuis un modèle du monde (UBM). Cette chaîne de traitement GMM-UBM est détaillée ci-après.

#### 2.1.3.1 Modélisation par GMM-UBM

**Modèle de mélange de Gaussiennes** Un GMM est un modèle statistique représentant un échantillon comme suivant une somme pondérée de $M$ lois gaussiennes. Sa densité s'écrit alors :

$$P(\mathbf{x}|\lambda_{aud}) = \sum_{i=1}^{M} w_i \mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i) \tag{2.8}$$

où :

$$\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_i|}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i (\mathbf{x}-\mu_i)} \tag{2.9}$$

et : $\sum_{i=1}^{M} w_i = 1$ avec $\mathbf{x}$ notre échantillon de $\mathbb{R}^D$ et $w_i$ le poids associé à chaque composante gaussienne de densité $\mathcal{N}(\mathbf{x}|\mu_i, \Sigma_i)$ ; où $\mu_i$ est sa moyenne et $\Sigma_i$ sa matrice de covariance.

Les paramètres du modèle à estimer sont l'ensemble des moments de chaque loi gaussienne (moyenne et matrice de covariance) ainsi que les poids à assigner à chaque composante, définissant ainsi la signature audio d'un locuteur :

$$\lambda_{aud} = \{\mu_i, \Sigma_i, w_i\}_{i=1}^{M} \tag{2.10}$$

Par souci de compacité et pour limiter le nombre de paramètres à estimer, nous considérons le plus souvent des matrices de covariance diagonales, et non pleines. L'estimation des paramètres est généralement opérée par estimation du maximum de vraisemblance (MLE, pour « Maximum Likelihood Estimation ») [Fis25], et en pratique réalisée via l'algorithme Espérance-Maximisation (EM, pour « Expectation-Maximisation ») [DLR77 ; Bis06].

**Adaptation d'un modèle du monde** Un problème majeur de l'apprentissage direct d'un GMM pour caractériser la signature d'un locuteur, et de l'apprentissage automatique en général, est la quantité de données nécessaire à l'estimation efficace du modèle. Dans de nombreux contextes, dont le notre, une modélisation rapide est désirée, de l'ordre de quelques minutes d'enregistrement sonore. L'apprentissage direct d'un GMM en nécessitant bien plus, [RQD00] suggère alors d'avoir recours à l'adaptation d'un modèle du monde (UBM pour « Universal Background Model ») aux données (limitées) du locuteur à modéliser. Nous entendons par modèle du monde un GMM appris en amont à partir de plusieurs centaines d'heures d'enregistrement, préférablement dans des conditions proche des données à traiter. Deux modèles peuvent également être générés, un par genre. L'UBM est ensuite adapté par la méthode de Maximum A Posteriori (MAP) :

$$\lambda_{aud} = \underset{\lambda}{\operatorname{argmax}}\, P(X|\lambda)P(\lambda) \tag{2.11}$$

Nous cherchons le meilleur ensemble de paramètres $\lambda_{aud}$ qui maximise la probabilité *a posteriori*, avec $X$ le vecteur de coefficients, $P(X|\lambda)$ la vraisemblance de $\lambda$, $P(\lambda)$ la probabilité *a priori* de $\lambda$. En pratique, l'adaptation des moyennes des gaussiennes est souvent suffisante à une

adaptation performante du modèle du monde, laissant les poids et les matrices de covariance inchangées.

L'étape de vérification du locuteur est enfin opérée par comparaison des log-vraisemblances d'appartenance à un modèle d'un locuteur *vs.* le modèle du monde. Soit $\mathbf{y}$ un vecteur de paramètres extraits d'un segment de parole, le rapport des vraisemblances en échelle logarithmique (LLR pour Log Likelihood Ratio) s'exprime comme suit :

$$LLR_i(\mathbf{y}) = log\left(P(\mathbf{y}|\lambda_i)\right) - log(P(\mathbf{y}|\lambda_{UBM})) \tag{2.12}$$

Ainsi :
— si $LLR_i(\mathbf{y}) > 0$, alors le segment $\mathbf{y}$ est classé comme prononcé par le locuteur $i$.
— si $LLR_i(\mathbf{y}) \leq 0, \forall i \in [1, \ldots, N]$, avec $N$ le nombre de signatures déjà apprises, alors le segment $\mathbf{y}$ est classé comme prononcé par un nouveau locuteur, dont la signature sera ajoutée au catalogue.

**Implémentation** La construction de la signature audio a été réalisée à l'aide de la plate-forme open-source ALIZE [Lar+13] et de sa boîte à outils LIA-RAL, ainsi que de la librairie SPro [10] pour l'extraction des paramètres cepstraux. Nous générons ainsi un modèle du monde composé de 512 lois gaussiennes, à matrices de covariances diagonales, à partir de données extraites du corpus BREF80 [Lam+] amputé de deux locuteurs, respectivement le locuteur testé et un imposteur. Ces derniers sont scindés en données d'apprentissage et données de test. Suivant le protocole expérimental décrit en préambule de ce chapitre, nous évaluons les performances de la reconnaissance de locuteur aux trois niveaux de bruit du corpus, pour des segments de parole émis à plusieurs distances du microphone. Les résultats, en terme de précision et de rappel, sont présentés dans le tableau 2.2.

Table 2.2 – Performances de la reconnaissance du locuteur à trois niveaux de bruits.

| | **RSB (dB)** | | |
|---|---|---|---|
| **Paramètres** | 13.3 | 6 | -3.4 |
| Rappel/Précision | 1.0/1.0 | 0.91/1.0 | 0.33/1.0 |

Si le score de précision n'est ici pas très significatif par le faible nombre de locuteurs testés, le score de rappel exhibe la sensibilité du système au RSB. De plus, aux trois valeurs de SNR, le LLR subit une dégradation avec l'augmentation de la distance source-microphone : la reconnaissance est d'autant plus fiable que la source est proche. Cette propriété sera exploitée dans la section 3.2.

**Conclusion** La chaîne de traitement illustrée sur le synoptique 2.3 a conduit à l'apprentissage d'une signature audio d'un locuteur à travers les étapes suivantes : détection d'activité vocale par exploitation de la modulation de l'énergie à 4 Hertz, extraction des MFCC sur les segments de parole et modélisation GMM par une adaptation MAP d'un modèle du monde. La signature audio est notée $\lambda_{aud}$ et composée des moments (moyenne, matrice de covariance) de chaque loi composant le mélange gaussien, ainsi que les poids associés.

Une démarche similaire est présentée ci-après pour apprendre la signature vidéo d'une personne.

10. http ://spro.gforge.inria.fr/

## 2.2 Signature Vidéo

Cette section est consacrée à l'étude de la chaîne de traitement utilisée pour construire une signature visuelle d'un individu à partir d'un flux vidéo. Cette chaîne est illustrée sur le diagramme 2.7. Nous en détaillons chacune des étapes, représentées par des blocs, ci-après.

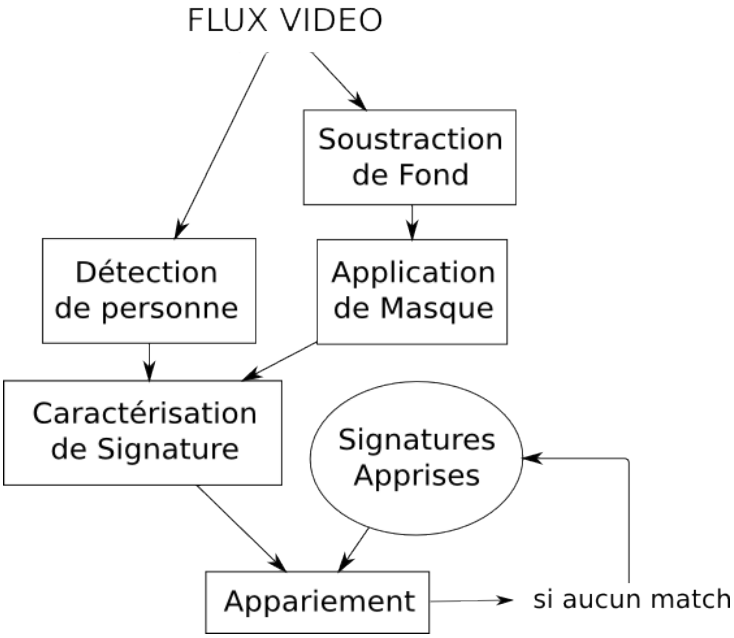



FIGURE 2.7 – Chaîne de traitement pour la génération d'une signature vidéo.

À partir du flux vidéo provenant d'une caméra ambiante, deux processus opèrent en parallèle : le premier est responsable de la détection de personnes dans l'image, le second de la soustraction de l'arrière-plan de l'image, afin d'isoler les pixels de l'image correspondant aux silhouettes des cibles. La section 2.2.1 sera consacrée au choix du détecteur à partir d'évaluations préalables de détecteurs existants. À partir des boîtes englobantes ainsi extraites et du masquage de l'arrière plan, des descripteurs visuels discriminant les individus entre eux sont ensuite extraits. Cette étape, représentée par le bloc « Caractérisation de signature », fait l'objet de la section 2.2.2. Enfin, la section 2.2.3 décrit la mise en correspondance de ces derniers ainsi qu'à l'apprentissage final de la signature.

### 2.2.1 Détection de personnes

La détection de personnes consiste à segmenter les régions contenant des personnes dans une image numérique. De même que pour un détecteur d'objet, dont il hérite, un détecteur de personnes délivre l'ensemble des hypothèses de présence des cibles sous la forme suivante :

$$P_i : \{x_i, y_i, l_i, h_i\}, \quad i = 1, \ldots, N \tag{2.13}$$

avec $P_i$ la $i$-ème personne parmi les $N$ détectées dans l'image, $(x_i, y_i)$ les coordonnées pixels du coin haut gauche de la boîte englobante, $l_i$ sa largeur et $h_i$ sa hauteur. Une illustration de détection (boîte englobante) est illustrée en figure 2.8.

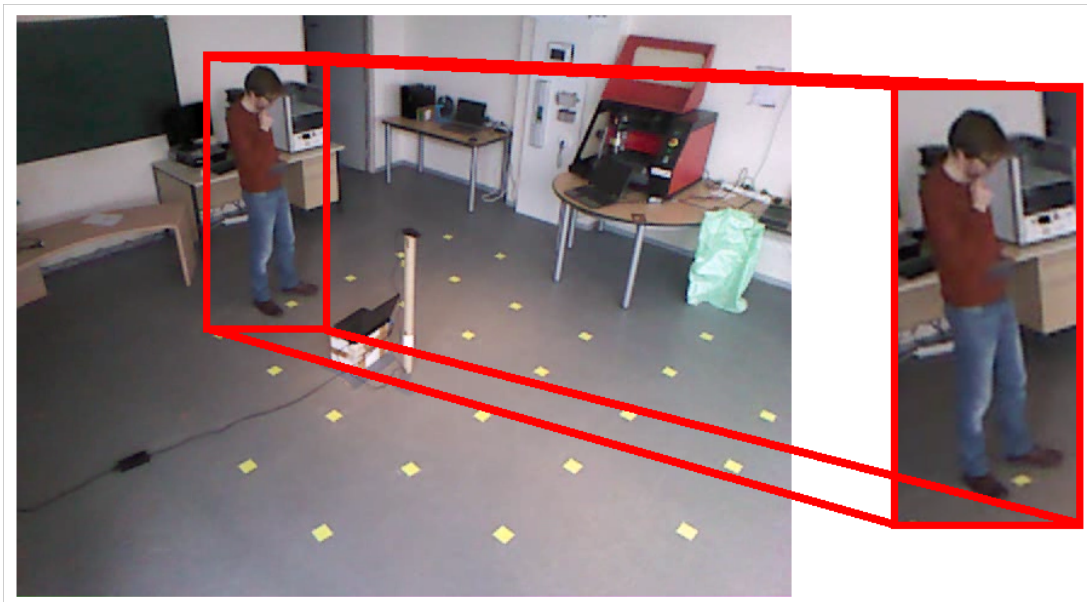


FIGURE 2.8 – Exemple de détection de personne sur notre corpus : extraction de la boîte englobante contenant la cible.

Elle trouve de nombreuses applications en surveillance, en robotique ou en sécurité routière, ce qui a conduit à l'implémentation de nombreux détecteurs ces vingt dernières années. Nous listons ci-après les détecteurs usuels et adaptés à notre problématique sachant que la littérature est riche sur le sujet. Le lecteur pourra se référer à [Zha+16] pour une synthèse des détecteurs existants.

#### 2.2.1.1 Focus sur les détecteurs visuels

Nous avons privilégié 4 détecteurs (HOG+SVM, DPM, ACF, RCNN), parmi les plus populaires, présentant un aperçu de la diversité des approches. Ils sont explicités dans les paragraphes suivants.

**HOG+SVM (Histogram of Oriented Gradients)** Dalal et Triggs proposent de caractériser l'apparence par des Histogrammes de Gradient Orienté, calculés localement sur des cellules carrées de quelques dizaines de pixels, formant une grille dense [DT05]. Cet ensemble d'histogrammes modélise l'orientation des contours d'un objet et se révèle très discriminant, ce qui en fait un descripteur de référence en détection. Une simple Machine à Vecteurs de Support (SVM) est ensuite entraînée pour la tâche de classification.

**DPM (Deformable Part Models)** À l'inverse des autres détecteurs présentés ici, le DPM opère sur des parties du corps séparément [Fel+10]. Ainsi des modèles à bases de variantes de HOG et de SVM latents sont appris pour chaque division, obtenant chacun un score. La présence d'une personne est ensuite déterminée par le rassemblement des décisions partielles.

**ACF (Aggregate Channel Features)** Ce détecteur a montré d'excellentes performances sur plusieurs bases de données publiques, pour un faible temps de calcul [Dol+14]. De l'image

sont extraits plusieurs canaux, pouvant contenir des informations de couleur ainsi que des variantes des HOG. Chaque canal est ensuite décomposé en blocs et les agrégats (la somme de ces blocs) sont regroupés en un vecteur, formant le descripteur. La décision est ensuite prise par un classifieur en cascade souple.

**RCNN (Regions with CNN Features)** Girshick extrait les régions d'intérêt probable de l'image, puis les injecte dans un réseau de neurones convolutionnel (CNN). Enfin la tâche de classification sur chaque classe du réseau est accomplie par des SVM [Gir+13]. Le souci majeur de cette méthode est sa lenteur, ainsi des variantes ont été proposées : le Fast-RCNN [Gir15] et le Faster-RCNN [Ren+15].

#### 2.2.1.2 Évaluation des détecteurs sur les bases de données ETH, CAVIAR et PETS

Les données visuelles que nous traitons sont bien moins marginales par rapport à la littérature que le sont nos données sonores. En effet, la composante vidéo de notre corpus présente des similarités avec les bases de données publiques présentées en chapitre 1. De plus, le nombre limité de participants et l'environnement d'acquisition non contraignant (espace ouvert, occultations rares) génère un jeu de données assez peu problématique. Ainsi nous évaluerons les méthodes de détection et de ré-identification sur des bases de données publiques afin d'éviter que la trop grande simplicité de nos données ne fausse les interprétations. Les performances de ces 4 détecteurs sont ainsi évaluées, en terme de précision et de rappel, sur les bases de données ETH, CAVIAR et PETS, régulièrement utilisées dans le MOT Challenge [11]. Les résultats sont synthétisés dans le tableau 2.3

Table 2.3 – Performances des détecteurs de personnes de l'état de l'art

| | **Détecteur (rappel/précision)** | | | |
|---|---|---|---|---|
| **Dataset** | HOG+SVM | DPM | ACF | RCNN |
| CAVIAR-EnterExit | 0.62/0.84 | 0.69/0.89 | 0.69/**0.94** | **0.73**/0.91 |
| CAVIAR-OneShop | 0.37/0.76 | 0.51/0.88 | 0.43/**0.95** | **0.57**/0.82 |
| PETS | 0.81/0.89 | 0.85/**0.95** | **0.92**/0.94 | 0.70/0.79 |
| ETH-Bahnoff | 0.52/0.47 | 0.57/0.73 | **0.63**/**0.78** | 0.58/0.75 |
| ETH-Jelmoli | 0.43/0.52 | **0.56**/0.79 | 0.52/**0.87** | 0.52/0.76 |
| ETH-Sunnyday | 0.67/0.60 | **0.76**/0.85 | 0.62/**0.91** | 0.69/0.81 |

Le détecteur ACF surclasse quasi-systématiquement ses concurrents, particulièrement en terme de précision ; exception faite sur la base de données PETS qui contient le plus grand nombre d'occlusions, mieux gérées par le DPM qui opère par parties, et non sur le corps complet. Le détecteur RCNN exploite pertinemment les possibilités offertes par les récentes avancées de l'apprentissage profond mais reste encore très coûteux et l'approche HOG+SVM, malgré des performances moindres, se distingue par sa simplicité algorithmique et son adéquation à des contextes simples. Le choix de notre détecteur de personnes se tournera alors vers l'ACF.

Pour la boîte englobante ainsi segmentée, nous extrayons ci-après des descripteurs discriminants qui constitueront la signature visuelle de la personne détectée.

11. `https://motchallenge.net/`

### 2.2.2 Représentation : problème de la ré-identification

Un état de l'art sur le problème de la ré-identification a été réalisé dans le chapitre 1. Nous justifions et détaillons dans cette partie le descripteur choisi pour construire notre signature visuelle.

#### 2.2.2.1 Descripteur SDALF (Symmetry-Driven Accumulation of Local Features)

Les travaux de Farenzena ont abouti à un modèle d'apparence ne nécessitant pas de technique d'apprentissage ni pour l'extraction des paramètres, ni pour leur mise en correspondance tout en atteignant des performances similaires à celles de méthodes supervisées [Far+10]. Il s'appuie sur une partition dynamique de la silhouette par la recherche d'axes de symétrie et d'antisymétrie ; partition sur laquelle sont extraits des informations de couleur et de texture. Le choix de ce descripteur est motivé non seulement par son procédé non supervisé, mais également par sa robustesse aux variabilités de point de vue, voire d'illumination, présentes dans notre contexte applicatif.

La partition dynamique de la silhouette et les différents descripteurs extraits sont explicités dans les paragraphes suivants.

**Partition de la silhouette** Plutôt que de scinder la boîte englobante fournie par le détecteur de personnes en considérant des proportions statiques, Farenzena propose de rechercher automatiquement des axes de symétrie et d'antisymétrie pour distinguer les différentes parties du corps, à savoir la tête, le torse et les jambes. Pour cela, il définit deux opérateurs. Le premier est l'Opérateur Chromatique Bilatéral :

$$C(i,\delta) = \sum_{B_{[i-\delta,i+\delta]}} d^2(p_i, \hat{p}_i) \tag{2.14}$$

où $d(.,.)$ est la distance euclidienne entre les valeurs HSV des pixels $p_i$ et $\hat{p}_i$, symétriques l'un de l'autre par rapport à l'axe horizontal situé à la hauteur $i$. Ces distances sont sommées sur une région de largeur $2\delta$ autour de l'axe horizontal en $i$. Cet opérateur sera maximisé lorsque l'axe situé en $i$ se confondra avec une transition de couleur, par exemple à l'intersection du buste et des jambes. Le deuxième opérateur est l'Opérateur de Couverture Spatiale, qui étudie les répartitions des pixels appartenant au premier plan. Afin d'obtenir les masques de premier plan pour chaque image, nous procéderons à une soustraction du fond, par un clustering GMM en ligne de l'arrière plan, comme détaillé en [Ziv04]. L'opérateur est défini ainsi :

$$S(i,\delta) = \frac{1}{J\delta} |A(B_{[i-\delta,i]}) - B_{[i,i+\delta]}| \tag{2.15}$$

avec $A(B_{[i-\delta,i]})$ l'aire du premier plan dans la région de largeur $J$ et délimité par les axes situé en $i-\delta$ et $i$.

Les axes de symétrie et d'antisymétrie peuvent être obtenus à partir de ces opérateurs. L'axe horizontal principal d'antisymétrie, séparant généralement le torse des jambes, est situé à la hauteur $i_{TL}$ :

$$i_{TL} = \underset{i}{\operatorname{argmin}}((1 - C(i,\delta)) + S(i,\delta)) \tag{2.16}$$

Le deuxième axe horizontal, cherchant les différences d'aires de pixels de premier plan dans la région inférieurement bornée par $i_{TL}$ , sépare la tête du torse, à la hauteur $i_{HT}$ :

$$i_{HT} = \underset{i}{\operatorname{argmin}}(-S(i,\delta)) \tag{2.17}$$

Enfin les axes de symétrie verticaux dans les deux régions $R_1$ et $R_2$, contenant respectivement le torse et les jambes, sont localisés en $j_{LRk}$, $k = 1, 2$ :

$$j_{LRk} = \underset{j}{\operatorname{argmin}}(C(j,\delta) + S(j,\delta)) \tag{2.18}$$

La création de la partition est illustrée sur la figure 2.9. Nous détaillons ensuite le calcul des 3 composantes du descripteur SDALF, respectivement les histogrammes HSV, les MSCR, et les RHCP.

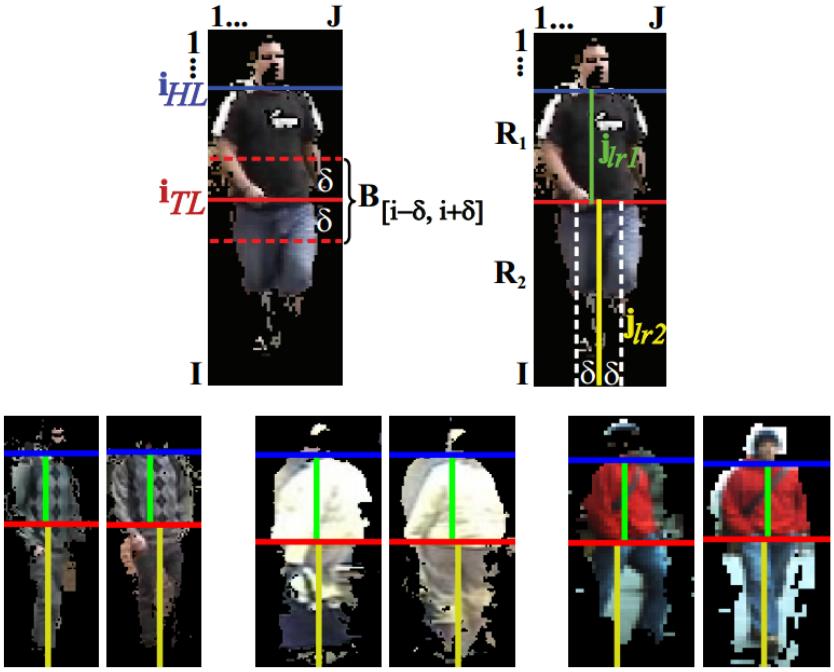



FIGURE 2.9 – Génération des axes de symétrie et d'antisymétrie, et exemples de partitions de plusieurs silhouettes. Image extraite de [Far+10].

**Histogrammes HSV (WH) :** L'espace de couleur HSV (Hue Saturation Value) est robuste aux changements de luminosité par séparation de l'information de couleur et d'intensité. La répartition des couleurs de chaque partie de la silhouette est ainsi représentée par un histogramme dans cet espace. Afin de renforcer le poids de l'information pertinente, chaque pixel est pondéré par un noyau gaussien centré sur l'axe de symétrie vertical correspondant. Nous évitons ainsi d'intégrer les résidus éventuels de l'arrière plan dans le calcul des histogrammes.

**Maximally Stable Color Region (MSCR) :** Forssen propose un descripteur qui évalue des distances de couleur entre des pixels d'une image pour en trouver les régions homogènes,

alors modélisées par des ellipses [For07]. Ces dernières sont ensuite représentées par leur aire, centroïde, matrice de second moment et leur couleur.

**Recurrent High-Structured Patches (RHSP) :** Ce paramètre, directement proposé par Farenzena, caractérise l'information de texture dans des zones à haute entropie en analysant les invariances de portions d'images tirées aléatoirement. De même que pour les histogrammes HSV, les portions seront tirées eu égard aux axes de symétrie, en réutilisant le noyau gaussien, favorisant l'information proche du centre de la silhouette. La récurrence d'un patch est évaluée par corrélation croisée locale.

Un exemple des composantes du SDALF extraits d'une paire d'images est illustré sur la figure 2.10

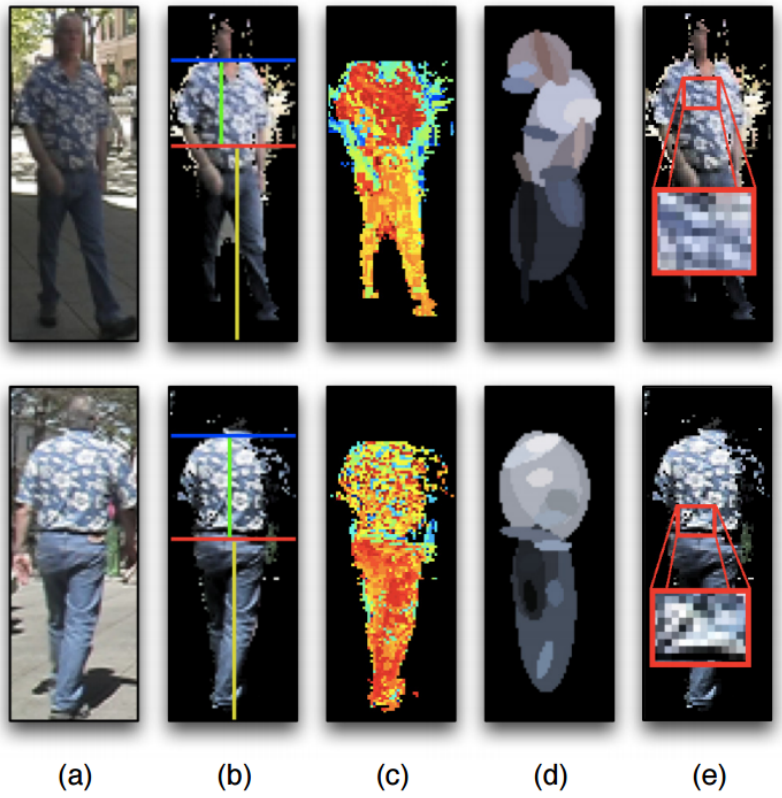



FIGURE 2.10 – Paramètres extraits de SDALF : en (a) une paire d'imagettes du même individu, en (b) les axes de symétrie et d'antisymétrie extraits, en (c) les histogrammes HSV, en (d) les MSCR et en (e) les RHCP. Image extraite de [Far+10].

### 2.2.3 Modèle et appariement de signatures

La décision d'un système de ré-identification revient à associer deux caractérisations issues des boîtes englobantes $I_A$ et $I_B$ en minimisant une distance entre les deux. Chaque composante du SDALF étant de nature différente et ayant des dimensions et échelles distinctes, 3 distances seront alors utilisées : $d_{WH}$, $d_{MSCR}$ et $d_{RHCP}$.

La distance $d_{WH}$ emploie la distance de Bhattacharrya, calculée entre les histogrammes de chaque partie du corps [Bha46]. Pour $d_{MSCR}$, nous recherchons les distances les plus faibles parmi les combinaisons possibles des ellipses générées. Nous opérons à l'aide de distances euclidiennes entre leurs centroïdes, et entre leurs couleurs. $d_{RHCP}$ est calculée sur la meilleure paire de patches, également par la distance de Bhattacharrya sur des histogrammes des RHCP.

Nous évaluons ci-après les performances de chaque descripteur séparément ainsi que la combinaison des 3 selon l'implémentation originale de [Far+10].

#### 2.2.3.1 Performance des différents descripteurs

Le descripteur conçu par Farenzena devait pouvoir répondre à d'éventuels forts changements de luminosité, de variations de poses ou encore de différences de fonctions colorimétriques, dans le cas multi-caméras. Ils exploitent ainsi la complémentarité des trois descripteurs en utilisant une mesure de distance comme somme pondérée des 3 distances citées précédemment :

$$\begin{aligned} d(I_A, I_B) = \beta_{WH}.d_{WH}(WH(I_A), WH(I_B)) \\ + \beta_{MSCR}.d_{MSCR}(MSCR(I_A), MSCR(I_B)) \\ + \beta_{RHCP}.d_{RHCP}(RHCP(I_A), RHCP(I_B)) \end{aligned} \quad (2.19)$$

Utilisant des données de la base VIPER, les valeurs des poids ont été ajustées comme suit : $\beta_{WH} = 0.4$, $\beta_{MSCR} = 0.4$ et $\beta_{RHCP} = 0.2$, le descripteur complet gagnant ainsi en performance sur cette base de données.

Si la fusion de ces paramètres peut se révéler pertinente dans des contextes critiques, certains scenarii moins exigeants ne la nécessiteront pas. Nous évaluons la performance des descripteurs pris séparément ou combinés 2.19, sur 3 jeux de données : ETHZ1, ETHZ2 et ETHZ3, notamment moins complexes que VIPER et non limités à des uniques paires d'observation. Nos données étant extraites d'un flux vidéo, nous sommes dans le cas « Multi Shot », comme ETHZ. Les résultats sont illustrés sur la figure 2.11 et résumés dans le tableau 2.4.

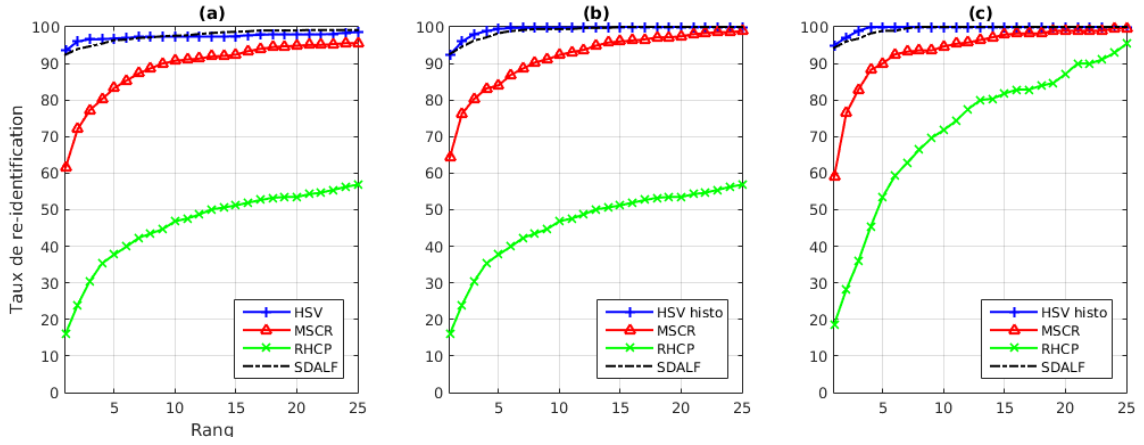



FIGURE 2.11 – Courbes CMC pour chaque composante séparée de SDALF sur les jeux de données ETHZ1 (a), ETHZ2 (b) et ETHZ3 (c).

Au vu des performances des histogrammes HSV par rapport à celles des MSCR et des RHCP, ces deux derniers ne seront plus pris en compte dans la construction de la signature

vidéo, gagnant ainsi en compacité, en coût calculatoire et en simplicité de représentation. De plus, sur la base ETHZ, le système de ré-identification n'employant que l'information apportée par les histogrammes HSV peut même devancer les scores de ré-identification obtenus par le descripteur SDALF complet. Ceci est dû aux pondérations $\beta_{WH}$, $\beta_{MSCR}$ et $\beta_{RHCP}$, apprises sur la base de données VIPER, et non optimales sur ETHZ. En ignorant la contribution des MSCR et des RHCP, outre le gain de compacité, nous nous affranchissons également du réglage des poids des paramètres. Au final : $\beta_{WH} = 1$, $\beta_{MSCR} = 0$ et $\beta_{RHCP} = 0$.

TABLE 2.4 – Score nAUC (normalized Area Under the Curve) pour les 3 paramètres du SDALF, ainsi que le descripteur complet, sur les trois jeux de données ETHZ1, ETHZ2 et ETHZ3.

| | **Dataset** | | |
|---|---|---|---|
| **Paramètres** | ETHZ1 | ETHZ2 | ETHZ3 |
| Histogrammes HSV | 99.02 | **99.49** | **99.68** |
| MSCR | 95.57 | 93.88 | 94.09 |
| RHCP | 68.23 | 68.23 | 74.42 |
| SDALF | **99.06** | 99.25 | 99.39 |

Enfin, afin de limiter les redondances, nous appliquons sur les histogrammes HSV de toutes les imagettes des individus un partitionnement en $k$-moyennes, et les $k$ histogrammes les plus proches des centroïdes constitueront la signature audiovisuelle, contenant ainsi les variations principales de vue et de pose. Le choix de $k$ dépendra de l'importance de ces variations (notamment dépendant de l'inclinaison de la caméra). La figure 2.12 illustre les imagettes sélectionnées après clustering pour $k$=6, sur notre corpus de 3 participants, $k$ étant fixé empiriquement.


FIGURE 2.12 – Images correspondante aux signatures vidéo des 3 personnes cibles.

Le faible nombre de participants ainsi que leur faible ressemblance résulte d'un taux de ré-identification de 100% sur ces acquisitions. Pour confronter une nouvelle observation, nous

calculerons la distance moyenne aux descripteurs des $k$ images qui constituent la signature. La signature vidéo d'un individu s'exprimera alors ainsi :

$$\lambda_{vid} = \{WH_i\}, \qquad i = 1, \ldots, k. \tag{2.20}$$

avec $k$ le nombre de clusters choisis.

**Synthèse** À travers les différentes évaluations réalisées, nous avons identifié les outils qui composent la chaîne de traitement illustrée sur le synoptique 2.7. La détection de personne est réalisée par la méthode ACF (Aggregate Channel Features), la description de l'apparence vestimentaire de la cible est ensuite générée par des histogrammes HSV pondérés par les axes de symétrie de la cible, et la similarité de deux apparences est évaluée par la distance de Bhattacharyya. La signature vidéo est enfin générée par un regroupement $k$-means des images de la cible.

## Conclusion

Ce chapitre a présenté une méthode d'apprentissage en ligne d'une signature audio $\lambda_{aud}$ et d'une signature vidéo $\lambda_{vid}$ d'une cible tout en justifiant les choix sous-jacents eu égard à la littérature. Les caractéristiques que nous avons modélisé sont le timbre du locuteur et la distribution colorimétrique sur des bandes corporelles relatives à l'anatomie humaine. Les blocs de la partie bleue du schéma-bloc 2.2 sont résumés dans le tableau 2.5.

Table 2.5 – Outils pour l'apprentissage des signatures audio et vidéo.

| Tâche | Outil |
|---|---|
| **Signature audio** | |
| Détection d'activité vocale | Modulation d'énergie à 4Hz |
| Descripteurs | 12 MFCC + énergie et leurs dérivées premières et secondes |
| Appariement et modèle | GMM-UBM et différence des log-vraisemblances |
| **Signature vidéo** | |
| Détection de personne | Aggregate Channel Features (ACF) |
| Descripteurs | Histogrammes HSV pondérés par les axes de symétrie |
| Appariement et modèle | Distance de Bhattacharrya et $k$-means |

Nous avons validé les choix de chacun des outils présentés ici à travers des évaluations réalisées sur des bases de données vidéo publiques proches de notre contexte applicatif, ainsi que sur un corpus personnel, pour les données audio.

Afin d'obtenir la signature audiovisuelle d'une personne cible, nous nous intéresserons dans le chapitre suivant à l'association, à travers deux nouvelles stratégies introduites, des observations audio et vidéo correspondantes.

# Chapitre 3

# Fusion par localisation des observations audio et vidéo

**Sommaire**



## Introduction

Dans le chapitre précédent, nous avons présenté deux systèmes indépendants d'apprentissage de signatures d'une personne, une audio notée $\lambda_{aud}$ et une vidéo notée $\lambda_{vid}$, à partir des percepts audio et vidéo respectivement fournis par un microphone et une caméra. L'objectif de ce chapitre est de fusionner ces signatures en une signature audiovisuelle d'une personne : $\lambda_{av}$.

Comme annoncé en introduction, la fusion des signatures audio et vidéo est réalisée par l'association d'observations hétérogènes issues de la même cible. Elle s'opère lorsque les détections audio et vidéo correspondantes seront cohérentes et compatibles spatialement et nécessite alors un processus de localisation.

**Localisation visuelle de la personne cible** Le plan du sol est calibré relativement à la caméra à l'aide des marqueurs placés sur le sol. Les projections des détections visuelles dans ce plan peuvent alors être obtenues en extrayant la position des pieds du sujet. Considérant la boîte englobante de hauteur $J$, nous l'estimerons à l'intersection de l'axe de symétrie $j_{LR2}$ : voir

la figure 2.9 et l'axe de hauteur $J/10$. Elle est représentée sur la figure 3.1 (a) par un cercle jaune.

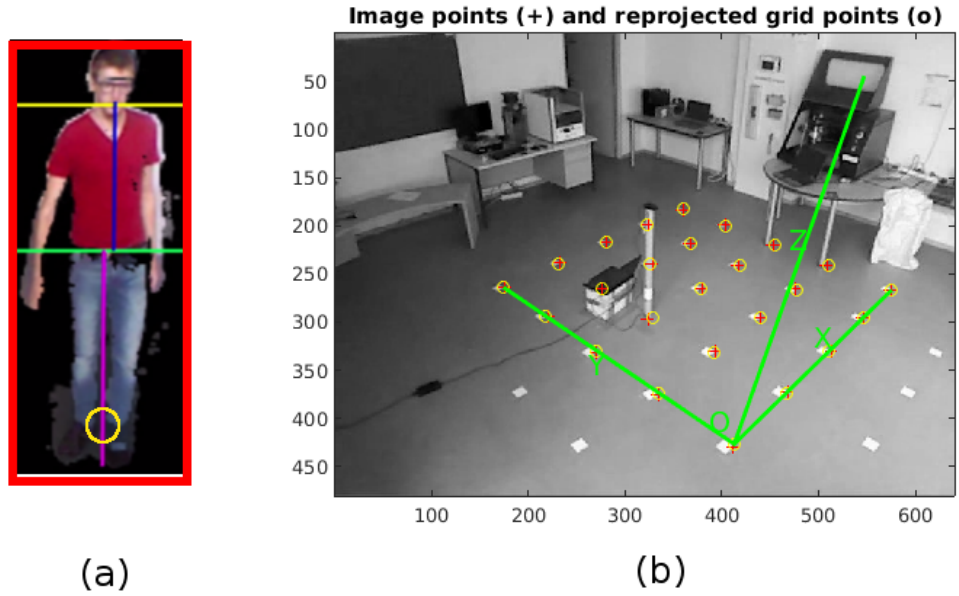



FIGURE 3.1 – Calibrage pour la localisation des observation vidéo. En (a) l'extraction dans le plan image de la position des pieds de la cible détectée, et en (b) la projection dans le plan image de la grille du repère caméra obtenu par calibration.

À l'image de la perception humaine, si les positions des détections vidéo peuvent être inférées aisément, la localisation des sources sonores se révèle bien plus délicate, et encore plus particulièrement dans le cas monocanal. Nous proposons dans ce chapitre deux méthodes de localisation grossière des observations audio. La première approche, détaillée en 3.1 vise à segmenter une zone de saillance audiovisuelle autour d'un microphone par l'étude du taux de réverbération dans le signal de parole. Dans cette zone les observations audio et vidéo détectées au même instant peuvent être ainsi associées. L'étendue des zones étant limitée, nous explorons diverses associations d'indices acoustiques présentant des corrélations à la distance source-microphone en section 3.3. Nous apprenons ainsi un estimateur de la distance source-microphone à plusieurs niveaux de bruit ambiant.

## 3.1 Localisation audio et limites observées

### 3.1.1 Stratégies existantes et positionnement de notre approche

Localiser une source sonore consiste à estimer ses coordonnées relatives par rapport au système d'écoute. À l'instar du système auditif humain, nous rechercherons principalement l'azimut, soit la direction de la provenance de la source (DOA, pour « Direction of Arrival »), souvent suffisant pour distinguer le locuteur actif.

#### 3.1.1.1 Paradigmes en localisation : binaural et traitement d'antennes

Deux stratégies ont alors émergé face à la problématique de localisation de source sonore : les approches binaurales et des approches fondées sur le traitement d'antenne. Ces dernières exploitent des réseaux de microphones en diverses configurations, linéaires, circulaires (pour estimer le DOA) ou rectangulaires (estimation du DOA et de l'élévation). Le traitement des signaux extraits des réseaux relèvent alors de l'estimation paramétrique et nous chercherons à maximiser la puissance reçue en fonction de l'angle d'arrivée à l'aide d'algorithmes comme MUSIC [Sch86], Root-MUSIC [RH89] ou ESPRIT [RPK86]. Les approches binaurales, elles, perdent la redondance d'information apportée par les réseaux, mais offrent un parallèle intéressant avec la perception humaine. Des indices comme la différence de niveaux inter-auraux, la différence de temps inter-auraux ou encore la différence de phase peuvent fournir une estimation de l'azimut. Enfin, la perception de l'élévation (axe bas-haut) est principalement effectuée à l'aide du pavillon de l'oreille qui filtre le son différemment en fonction de son angle d'incidence. Cette propriété est exploitée dans [SN09] où différents pavillons artificiels sont utilisés pour inférer le DOA à l'aide d'une unique source sonore.

Le cadre de cette thèse ne prévoit cependant pas d'utiliser de tels dispositifs, mais d'extraire de l'information à partir de microphones épars et donc sans recouvrement spatial. Il est donc nécessaire de repenser le problème de localisation, insoluble en l'état.

#### 3.1.1.2 Vers une approche monorale par estimation de distance

En plus de traiter la source sonore, il est possible d'extraire des indices spatiaux à partir de la résonance de la pièce. En environnement clos, un système de capture sonore va non seulement enregistrer les ondes directement propagées depuis la source, mais également les ondes réfléchies par les parois de la pièce d'acquisition. En réponse à une impulsion de Dirac, le signal acquis s'exprimera comme une somme d'une composante directe et de plusieurs réflexions d'intensité décroissante, formant la réverbération, en fonction des coefficients d'absorption des parois sur lesquelles l'onde s'est réfléchie. Cette relation entre la réverbération et l'aire de la pièce ainsi que les coefficients d'absorption de ses parois a été mise en évidence par Sabine en 1898. Elle s'exprime à l'aide du temps de réverbération, noté $T_{60}$ qui représente la durée que met la puissance d'une impulsion à décroître de 60 dB. L'estimation de cette mesure peut être effectuée de manière supervisée, en mesurant le déclin à l'arrêt d'un bruit blanc,, ou encore à partir de la réponse impulsionnelle de la salle [Sch65]. Nous pouvons aussi l'estimer de manière aveugle, en analysant les distributions du déclin de la parole dans le domaine spectral [WHN08] ou par des méthodes basées sur l'estimation du maximum de vraisemblance d'un modèle de déclin [Rat+03 ; RJO04]. Cependant ces méthodes peuvent nécessiter l'apprentissage au préalable de modèles, ou encore des coûts de calcul élevés. Une mesure d'intelligibilité de la parole a été proposée en [FZC10], exploitant le blanchissement du spectre de modulation du signal de parole en environnement réverbéré. La corrélation entre cette mesure et le $T_{60}$ a été mise en évidence dans [Gau+12]. Le calcul de cette mesure est détaillée dans le paragraphe suivant.

**Speech to Reverberation Modulation energy Ratio - SRMR** Comme évoqué précédemment, le spectre de modulation d'un signal de parole propre atteint son maximum vers 4 Hertz, correspondant au débit syllabique, et l'ensemble de ses composantes se situent dans les basses fréquences, typiquement dans la bande [2Hz-16Hz]. En revanche celui d'un signal de parole réverbéré s'étalera sur tout l'espace des fréquences de modulation [DFP94]. En effet, un signal réverbéré peut être exprimé comme la convolution entre le signal propre $x$ et la réponse impul-

sionnelle $r$ de la salle dans laquelle le son se propage, dont le comportement est proche d'un bruit blanc amorti par une enveloppe exponentielle :

$$s(n) = x(n) * r(n) \tag{3.1}$$

Le signal est au préalable passé dans un banc de 23 filtres gammatone, modélisation proche du traitement du son dans la cochlée. Les enveloppes $e_j$ des signaux résultants $s_j$ du filtrage par le $j$-ième canal, sont ensuite extraites à l'aide de la transformée de Hilbert $\mathcal{H}$ (Eq. 3.2), découpées en trames de 256 ms, puis le spectre de modulation est estimé à l'aide de la transformée de Fourier discrète (Eq. 3.3).

$$e_j(n) = \sqrt{s_j(n)^2 + \mathcal{H}\{s_j(n)^2\}} \tag{3.2}$$

$$E_j(i, f) = |DFT(e_j(i, k))|^2 \tag{3.3}$$

où $i$ et $k$ représentent les index respectifs de la trame et de l'échantillon dans la trame.

Le spectre de modulation ainsi obtenu est scindé en 8 bandes fréquentielles à l'aide d'un banc de filtres, inspiré par la perception auditive, dont les fréquences centrales et les largeurs de bande sont listées dans le tableau 3.1 :

Table 3.1 – Fréquences de modulation centrales ($f_c$) et bandes passantes ($BP$), en Hz, du banc de filtres

| | **Index de Bande** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $f_c(Hz)$ | 4.0 | 6.5 | 10.7 | 17.6 | 28.9 | 47.5 | 78.1 | 128.0 |
| $BP(Hz)$ | 1.9 | 3.4 | 5.9 | 9.8 | 15.9 | 26.4 | 43.2 | 70.8 |

Nous notons $\bar{\mathcal{E}}_{j,k}$ l'énergie de modulation moyenne sur l'ensemble des trames dans la bande de modulation $k$ et dans la bande $j$ du filtre gammatone, et $\bar{\mathcal{E}}_k$ l'énergie de modulation moyenne sur l'ensemble des trames et l'ensemble des bandes du banc de filtre gammatone, pour la bande de modulation $k$. La Figure 3.2 illustre le comportement de $\bar{\mathcal{E}}_k$ pour 4 niveaux de réverbération sur le même signal de parole.

Nous pouvons observer une nette inversion des rapports des énergies de modulation par bande au taux de réverbération entre les 4 premières bandes et les 4 dernières. Nous définissons alors le SRMR comme le rapport des premières bandes de modulation sur les dernières :

$$\mathrm{S}RMR = \frac{\sum_{k=1}^{4} \bar{\mathcal{E}}_k}{\sum_{k=5}^{8} \bar{\mathcal{E}}_k} \tag{3.4}$$

### 3.1.2 Taux de réverbération et indice de proximité

Cette partie traite de l'exploitation de la relation entre le taux de réverbération et la proximité source-microphone.

Plus la cible est proche du système d'écoute, plus le signal direct reçu est haut en énergie, la pression sonore subissant une perte de 6 dB en doublant la distance dans des espaces ouverts. Ainsi, éloigner une source sonore accroît la force de la réverbération du signal reçu. Afin de

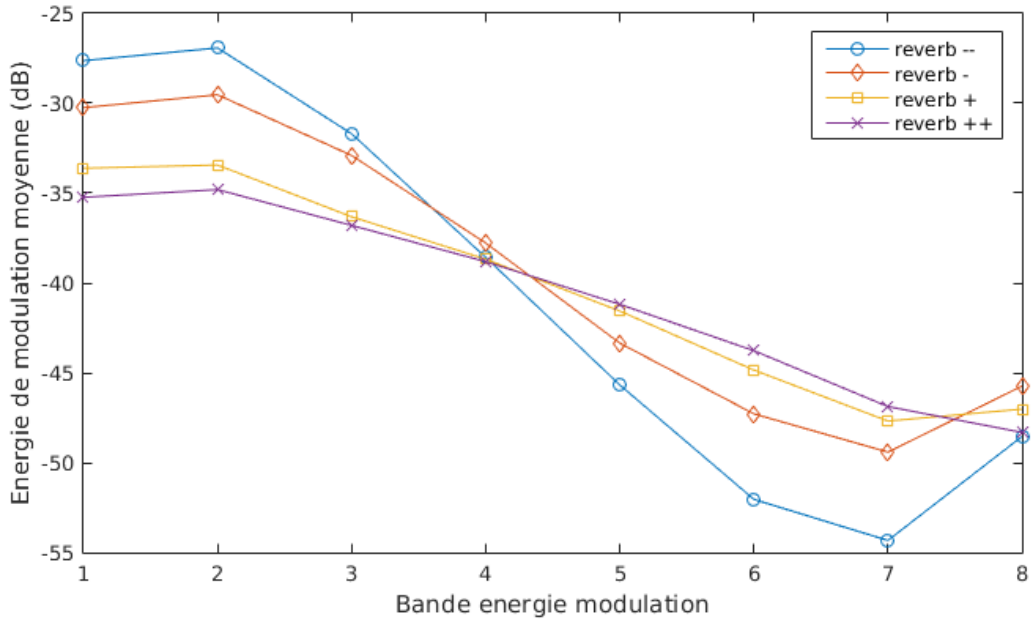



FIGURE 3.2 – Énergie de modulation par bande $\bar{\mathcal{E}}_k$ pour $k = 1, \ldots, 8$, à 4 niveaux de réverbération.

valider cette hypothèse, nous estimons le SRMR du même signal de parole émis à 3 distances différentes, en synthèse puis en conditions réelles. Le signal de référence, 16 secondes de parole lue en condition studio, est extrait du corpus BREF [Lam+]. Afin de synthétiser la distance, il est ensuite filtré par convolution avec 3 réponses impulsionnelles, issues de la base donnée MARDY [12], enregistrées respectivement à 1, 2 et 3 mètres de la source sonore. En conditions réelles, le même fichier de parole est diffusé à travers une enceinte à 1, 2 et 3 mètres, ainsi que juste à côté du microphone comme référence. Le SRMR est ensuite calculé sur des fenêtres d'une seconde sur chacun des signaux. Les résultats sont illustrés en figure 3.3. En (a) sont présentés les résultats de synthèse et en (b) les résultats en conditions réelles.

Les résultats en données réelles pour des distances proches, jusqu'à 1 mètre, corroborent l'hypothèse validée en données synthétiques de l'estimation du taux de réverbération comme indice de proximité. Cependant les limites du modèle sont atteintes ensuite, l'ensemble des réflexions restreignant la distinction des distances.

Nous étendons ces évaluations en couvrant cette fois l'ensemble de la salle d'acquisition selon le protocole expérimental décrit en préambule du chapitre 2. Par commodité le SRMR n'est pas calculé aux positions des microphones. Nous pouvons alors observer son évolution en fonction de la distance sur la figure 3.4. Deux comportements en ressortent : une dégradation pseudo-exponentielle dans une zone d'approximativement 1 mètre de rayon, centrée autour du microphone, et des valeurs plus faibles et plus homogènes dans le reste de la pièce.

Nous définissons l'Indice de Proximité Audio (IPA) suivant, à l'instant $t$ pour le micro $M_k$ :

$$IPA_{t,M_k} = \frac{1}{\alpha} SRMR_{t,M_k} \tag{3.5}$$

12. `http://www.commsp.ee.ic.ac.uk/~sap/resources/mardy-multichannel-acoustic-reverberation-database-at-york-database/`

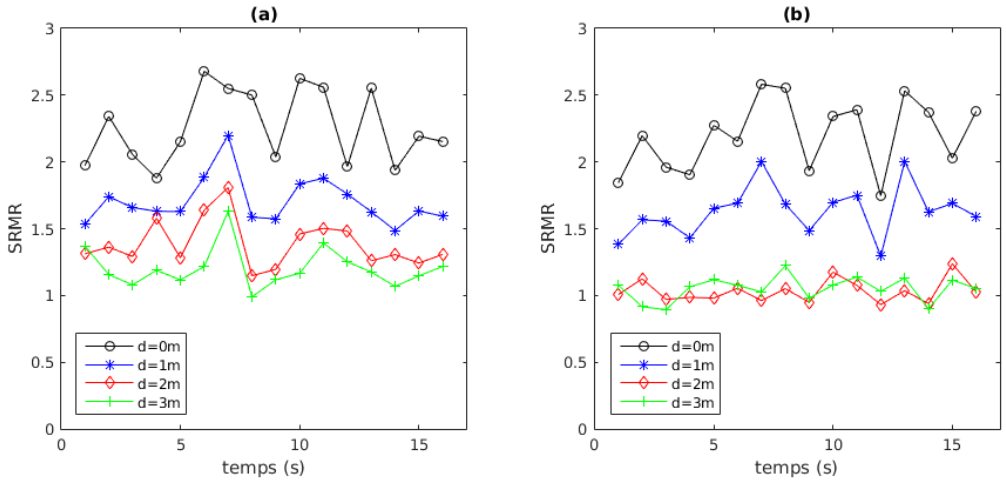



FIGURE 3.3 – SRMR sur un signal de parole émis à plusieurs distances, en synthèse (a) et en données réelles (b).

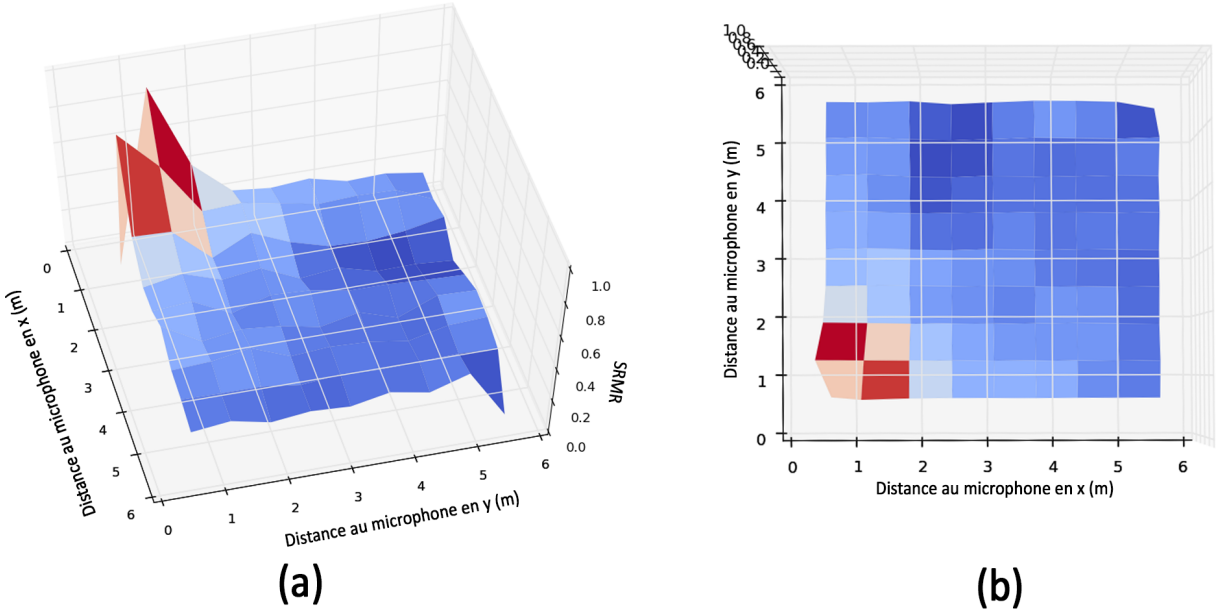



FIGURE 3.4 – Évolution du SRMR en fonction de la distance au microphone en vue 3D en (a) et zénithale en (b)

avec $\alpha$ coefficient de normalisation, fixé empiriquement, pour borner l'indice entre 0 et 1. Dans la section suivante l'intégration des estimations de proximité pour la fusion des signatures est traitée à travers deux approches : une approche par zones, en section 3.2.1 et une approche généralisée en section 3.3.

## 3.2 Fusion bimodale pour une signature audiovisuelle

Nous entendons par fusion l'agrégation de données, de descripteurs ou de scores, dans le but de perfectionner, d'enrichir un modèle ou une estimation. Elle peut avoir lieu à plusieurs niveaux, en fonction des corrélations présumées des entrées. De manière non exhaustive, nous pouvons en distinguer trois principales :

— fusion précoce : lorsque les données à traiter sont directement corrélées et présumées complémentaires,
— fusion intermédiaire : lorsque nous cherchons à aggréger des paramètres extraits de données, possiblement hétérogènes, visant à caractériser la même information,
— fusion tardive : aussi nommée *fusion de décision*, lorsqu'elle est effectuée sur la sortie (décisions binaires ou scores) de plusieurs systèmes de classification par exemple, utilisant des données et des paramètres n'ayant pas de lien apparent entre eux.

### 3.2.1 Stratégie n° 1 : fusion tardive

Par rapport à la classification précédente, l'association des signatures sera effectuée en fusion tardive. Deux observations hétérogènes, une audio et une vidéo, seront jointes lorsqu'elles proviennent de positions suffisamment proches. Cependant l'étude menée en section 3.1 montre que l'inférence de la position d'une source sonore est irréalisable dans notre contexte, et que seule la forte proximité peut être détectée avec une confiance acceptable. Nous exploiterons alors ce comportement proximal pour extraire des zones dans lesquelles seront fusionnées les signatures.

#### 3.2.1.1 Proximité et saillance audiovisuelle

Nous désignons par zone de saillance audiovisuelle l'ensemble des positions sur lesquelles un indice de proximité audiovisuel atteint des valeurs résolument distinctes dans l'espace d'acquisition. Elle sera centrée autour d'un microphone, noté $M_k$ et de position $(x_{M_k}, y_{M_k})$. La localisation de l'observation vidéo, notée $(x, y)$ dans le plan du sol étant elle réalisable, il est possible d'en extraire un Indice de Proximité Vidéo (IPV) comme l'inverse de la distance euclidienne dans le plan du sol entre la position estimée de l'observation et celle du microphone $M_k$ :

$$IPV_{t,M_k} = \frac{1}{\sqrt{(x - x_{M_k})^2 + (y - y_{M_k})^2}} \tag{3.6}$$

Enfin nous exprimerons l'Indice de Proximité Audio Vidéo comme le produit des deux indices, mesure de la saillance audiovisuelle pour deux observations hétérogènes, une sonore et une visuelle, à l'instant $t$ :

$$\text{IPAV}_{t,M_k} = IPA_{t,M_k} * IPV_{t,M_k} \tag{3.7}$$

La délimitation de la zone est fonction d'un seuil $th$, nécessitant un compromis entre sa largeur et sa robustesse.

#### 3.2.1.2 Segmentation des zones de saillance

Considérons le protocole expérimental décrit dans le préambule du chapitre 2, utilisant les percepts issus de la caméra 1 et du micro 2. Ce dernier étant placé au centre de la salle, nous pouvons espérer définir autour une zone circulaire. L'IPAV est calculé à chaque position, hormis celle où est placée le microphone, et les résultats sont illustrés en figure 3.5.

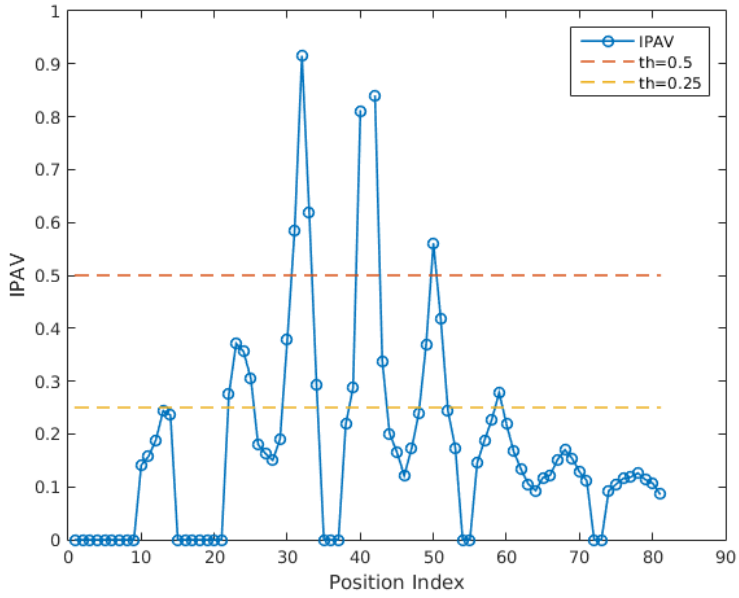



FIGURE 3.5 – Indice de Proximité Audio Vidéo calculé sur tout l'espace d'acquisition. Les maxima locaux correspondent aux positions voisines du microphone. À titre d'exemple, deux valeurs de $th$ sont affichées, dessinant des zones de largeurs différentes.

Afin d'optimiser la taille et la robustesse des zones, nous découpons nos données en un jeu d'apprentissage et un jeu de test, extraites de 3 sessions à 3 locuteurs différents. Nous faisons varier $th$ de 0,1 à 0,5 avec un pas de 0,05 et pour chaque valeur. Nous apprenons les contours de la zone de saillance grâce un SVM à noyau gaussien. La robustesse de la zone est quantifiée en terme de taux d'erreur de classification binaire (dans la zone *versus* hors zone) pour chaque locuteur. Les résultats sont illustrés sur la figure 3.6.

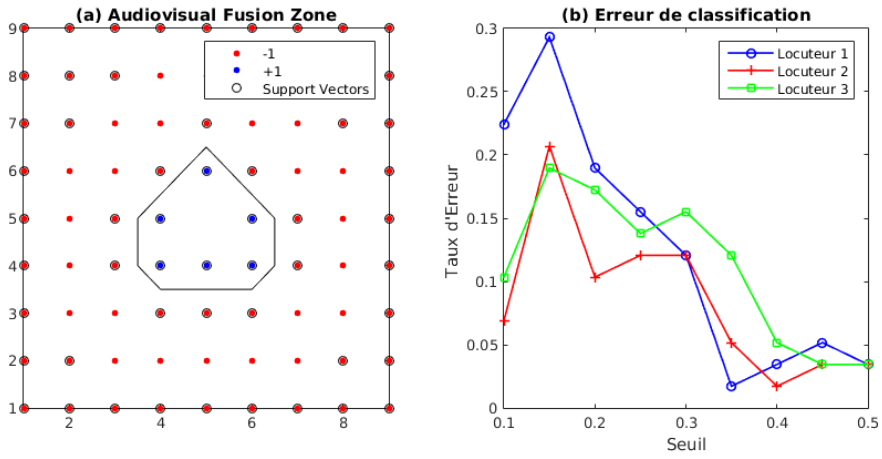



FIGURE 3.6 – Réglage du contour de la zone. En (a) la classification apprise par le SVM pour $th$=0.4, en (b) l'erreur de classification en fonction de $th$ pour les 3 locuteurs

Un compromis satisfaisant est atteint pour $th = 0,4$ pour les 3 locuteurs, délimitant des zones de largeur 1,3 m et une erreur de classification faible. La classification des observations est illustrée en figure 3.7.

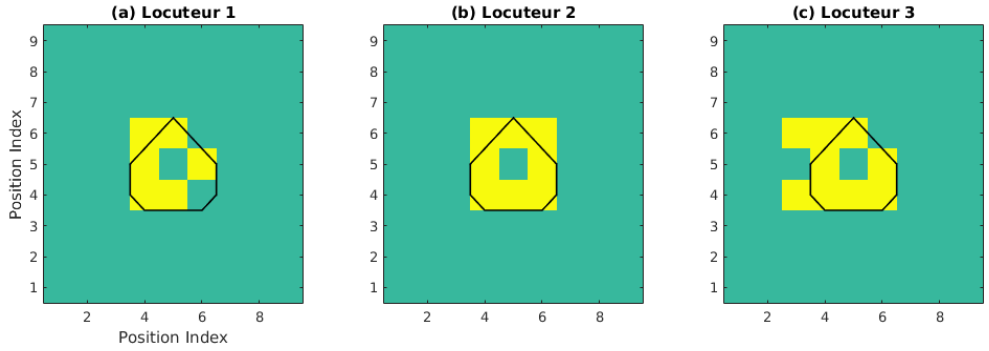



FIGURE 3.7 – Classification des positions à chaque position pour les 3 locuteurs. En jaune les positions observations classées saillantes, et en noir le contour de la zone de saillance. Erreurs de classification : locuteur 1 : 3,75%, locuteur 2 : 2,5%, locuteur 3 : 5%.

Cette méthode de fusion audiovisuelle pour la construction de la signature audio-vidéo d'un individu est publiée dans [Dec+16]. L'inconvénient principal de cette méthode est la restriction spatiale de la sensibilité audiovisuelle de la salle. La fusion des signatures audio et vidéo est contrainte par le comportement des usagers, peu en phase avec le contexte non supervisé de l'étude. En effet, un usager ne se déplaçant jamais dans la zone de saillance audiovisuelle définie rendra impossible la fusion. Cependant, le placement des microphones à certains emplacements stratégiques (bureau de l'enseignant, paillasse de Travaux Pratiques...) eu égard au contexte applicatif, permet de maximiser les chances d'associations audio-vidéo par l'exploitation de ses percepts.

Nous présentons ci-après une seconde stratégie de fusion des signatures audiovisuelles opérant sur un espace plus étendu.

## 3.3 Stratégie n° 2 : localisation de la cible

Le taux de réverbération dans un signal de parole se révèle être un estimateur relativement robuste de la forte proximité d'une source sonore à un microphone, mais il est rapidement mis en défaut à des distances plus importantes. Dans cette section nous nous attacherons à affiner cette estimation, non plus en définissant des zones saillantes, mais en tentant d'extraire une réelle information de distance à partir d'un simple flux audio monocanal. L'emploi de nouvelles mesures complémentaires est alors nécessaire.

### 3.3.1 Fusion de mesures audio pour l'estimation de la distance

Le choix d'exploiter le SRMR comme mesure est motivé par sa robustesse présumée aux variabilités intra et inter-locuteurs, à l'inverse d'autres paramètres, notamment l'énergie, dépendante de l'effort vocal, impossible à estimer en contexte de champ lointain. Cependant il est possible de considérer ses variations limitées, eu égard à la théorie de l'accommodation communicative [Gil] qui exhibe le phénomène de convergence au sein d'un groupe d'individus, soit l'altération de leurs paramètres non-verbaux comme leur prononciation, les temps et fréquences des pauses, le débit

syllabique ou particulièrement l'intensité vocale dans le but de les harmoniser. En apprenant nos modèles avec des données émises à une intensité vocale standard, nous espérons ainsi décrire une caractérisation moyenne avec une variance raisonnable. Enfin, outre l'énergie du signal et le SRMR, la log-vraisemblance de la reconnaissance du locuteur semble également présenter une dégradation proportionnelle à la distance de la source sonore et constitue un troisième paramètre pouvant porter l'information spatiale recherchée. Dans la suite de cette partie nous explorerons alors la fusion de ces paramètres pour l'estimation de la distance d'une source sonore.

Comme indiqué en introduction de cette section, nous déciderons du niveau d'une opération de fusion en fonction des corrélations supposées entre les données ou les paramètres considérés. À l'inverse de la fusion des signatures sonores et visuelles, totalement décorrélées l'une de l'autre, les trois paramètres ci-dessus sont tous dépendants de la distance au locuteur. Nous nous placerons alors dans un contexte de fusion intermédiaire, soit une fusion de paramètres dans le but d'exhiber leur information mutuelle.

#### 3.3.1.1 Un outil : l'Analyse Canonique des Corrélations

Les problèmes majeurs de la fusion de descripteurs sont les différences de dimensions et d'échelles de ceux-ci. Ces dernières nécessitent des opérations préalables sur les descripteurs pour les rendre comparables et conjointement traitables. Parmi les méthodes relevant de la statistique descriptive multidimensionnelle, la plus notable est l'Analyse en Composantes Principales (ACP) qui vise à réduire la redondance au sein de l'ensemble des variables par projection dans un espace de dimension réduite et conservant un maximum d'information. L'Analyse Factorielle des Correspondances (AFC) poursuit le même objectif tout en offrant un espace de représentation commun aux variables et aux individus. Dans notre contexte, nous avons cependant une connaissance *a priori* des corrélations des trois variables considérées, toutes dépendantes de la distance de la source sonore au microphone, et c'est cette information mutuelle que nous cherchons à cristalliser.

L'Analyse Canonique des Corrélations (ACC), introduite dans [Hot36], effectue une projection de deux vecteurs de variables $\mathbf{x}$ et $\mathbf{y}$ de dimensions respectives $N_x$ et $N_y$ dans un nouvel espace commun de dimension $N$ : $N \leq min(N_x, N_y)$ tel que les corrélations croisées entre ceux-ci soit maximales.

Les transformées ACC $\mathbf{x}'$ et $\mathbf{y}'$ s'exprimeront alors de la manière suivante :

$$\mathbf{x}' = \mathbf{H}_x \mathbf{x} \tag{3.8}$$

$$\mathbf{y}' = \mathbf{H}_y \mathbf{y} \tag{3.9}$$

où $\mathbf{H}_x$ et $\mathbf{H}_y$ décrivent les transformations linéaires projetant $\mathbf{x}$ et $\mathbf{y}$ dans l'espace de dimension $N$. Elles sont représentées par des matrices de tailles respectives $N \times N_x$ et $N \times N_y$ :

$$\mathbf{H}_x = \begin{bmatrix} \mathbf{h}_{x1}^T \\ \mathbf{h}_{x2}^T \\ \vdots \\ \mathbf{h}_{xN}^T \end{bmatrix}, \quad \mathbf{H}_y = \begin{bmatrix} \mathbf{h}_{y1}^T \\ \mathbf{h}_{y2}^T \\ \vdots \\ \mathbf{h}_{yN}^T \end{bmatrix} \tag{3.10}$$

Les lignes de ces matrices forment alors une base orthonormale pour le nouvel espace. Nous pouvons les calculer en résolvant le système d'équations aux valeurs propres suivant :

$$\begin{cases} \mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{h}_x = \gamma^2 \mathbf{h}_x \\ \mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{h}_y = \gamma^2 \mathbf{h}_y \end{cases} \tag{3.11}$$

où les vecteurs propres correspondent aux vecteurs de la base de l'espace ACC, $\gamma$ leurs valeurs propres associées et en notant $\mathbf{C}$ la matrice de covariance de $\mathbf{x}$ et $\mathbf{y}$ :

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{bmatrix} \tag{3.12}$$

Cette méthode a trouvé des applications dans plusieurs domaines [HSST04], et notamment en fusion audiovisuelle [Sar+07].

Dans notre contexte, plutôt que de comparer les paramètres entre eux, nous effectuerons une ACC entre la vérité terrain, soit l'inverse de la distance de la source sonore, et la concaténation de nos 3 paramètres calculés à chaque position de la source sonore. Ainsi, nous nous assurons de mettre en avant la dépendance à la distance dans chaque paramètre. En reprenant les notations précédentes nous avons alors :

$$\mathbf{x} = \begin{bmatrix} \frac{1}{\mathbf{d}} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \textbf{Énergie} \\ \textbf{SRMR} \\ \textbf{logV} \end{bmatrix} \tag{3.13}$$

avec $\mathbf{d}$ le vecteur contenant les distances euclidiennes source-microphone et $\mathbf{logV}$ le vecteur des log-vraisemblances en sortie du système de reconnaissance de locuteur. Pour apprendre les matrices $\mathbf{H}_x$ et $\mathbf{H}_y$, nous utilisons un jeu de données d'apprentissage générées en diffusant de la parole produite par un locuteur à 81 positions uniformément réparties dans la salle, formant un quadrillage, suivant le protocole décrit en préambule du chapitre 2. Les paramètres extraits, ainsi que la transformée ACC de ceux-ci sont illustrés sur la figure 3.8. Par commodité de visualisation les paramètres affichés en (a) sont ici normalisés par leur maximum.

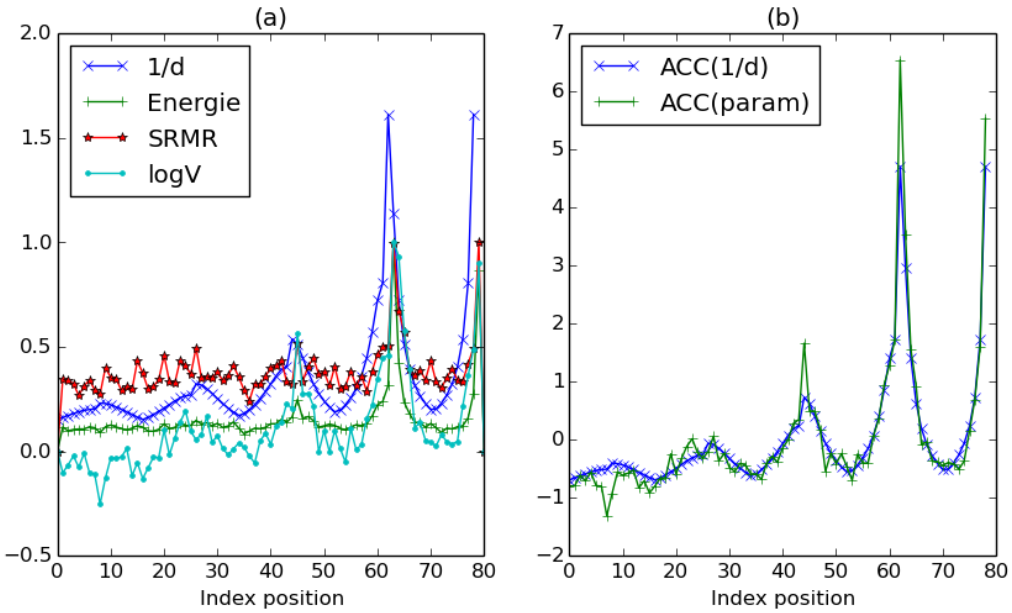



FIGURE 3.8 – ACC entre un vecteur de trois paramètres et l'inverse de la distance de la source sonore pour plusieurs positions, en (a) les paramètres, en (b) leur transformée par ACC.

**Évaluation des performances des combinaisons de paramètres** Les 3 paramètres considérés individuellement semblent être des estimateurs uniquement efficaces lorsque la distance source-microphone est faible (aux maxima locaux de la courbe $1/d$), mais un gain est observable en combinant les 3 par l'ACC. Nous le quantifions en calculant la corrélation de Pearson (équation 3.14) entre différentes combinaisons de paramètres sur 3 jeux de données plus ou moins bruitées.

$$r(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i-\overline{\mathbf{x}})(y_i-\overline{\mathbf{y}})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{\mathbf{y}})^2}} \tag{3.14}$$

Les résultats sont présentés dans le tableau 3.2. En confrontant les meilleurs scores extraits des combinaisons d'au moins 2 paramètres avec les meilleurs résultats issus des paramètres considérés individuellement, des gains de 4,8%, 4,1% et 0,4% sont obtenus pour les signaux respectivement propres, faiblement et fortement bruités, validant ainsi l'apport de la méthode de fusion.

TABLE 3.2 – Corrélation de Pearson entre différentes combinaisons de paramètres et l'inverse de la distance.

| | **RSB (dB)** | | |
|---|---|---|---|
| **Paramètres** | 13.3 | 6 | -3.4 |
| logV | 0.825 | 0.849 | 0.914 |
| SRMR | 0.85 | 0.934 | 0.938 |
| Énergie | 0.901 | 0.928 | 0.917 |
| Énergie + logV | 0.941 | 0.969 | 0.922 |
| Énergie + SRMR | 0.939 | 0.969 | 0.928 |
| SRMR + logV | 0.921 | 0.969 | **0.942** |
| Énergie + SRMR + logV | **0.949** | **0.975** | 0.937 |

Afin d'évaluer leur stationnarité, les performances des descripteurs sont testées sur un nouveau jeu de données issu de la même session expérimentale que celle qui a servi à générer les données d'apprentissage. Les 4 descripteurs, correspondant aux différentes combinaisons des 3 paramètres, sont calculés et projetés dans l'espace appris par l'ACC. Leurs valeurs sont illustrées sur la figure 3.9 et la qualité de l'estimation est évaluée en terme d'erreur quadratique moyenne (MSE), dont les valeurs sont exposées pour les 4 configurations dans le tableau 3.3 :

$$MSE(\mathbf{y}') = \frac{1}{n}\sum_{i=1}^{n}(x_i'-y_i')^2 \tag{3.15}$$

Les estimations utilisant les configurations (a), (c) et (d) possèdent comme paramètre commun la vraisemblance de la reconnaissance du locuteur (notation « logV ») et présentent toutes un décalage par rapport à la référence, les rendant inefficaces. Ce décalage est en revanche absent de la configuration (b), qui combine uniquement l'énergie et le SRMR et qui sera alors choisie comme estimateur de la distance source-microphone.

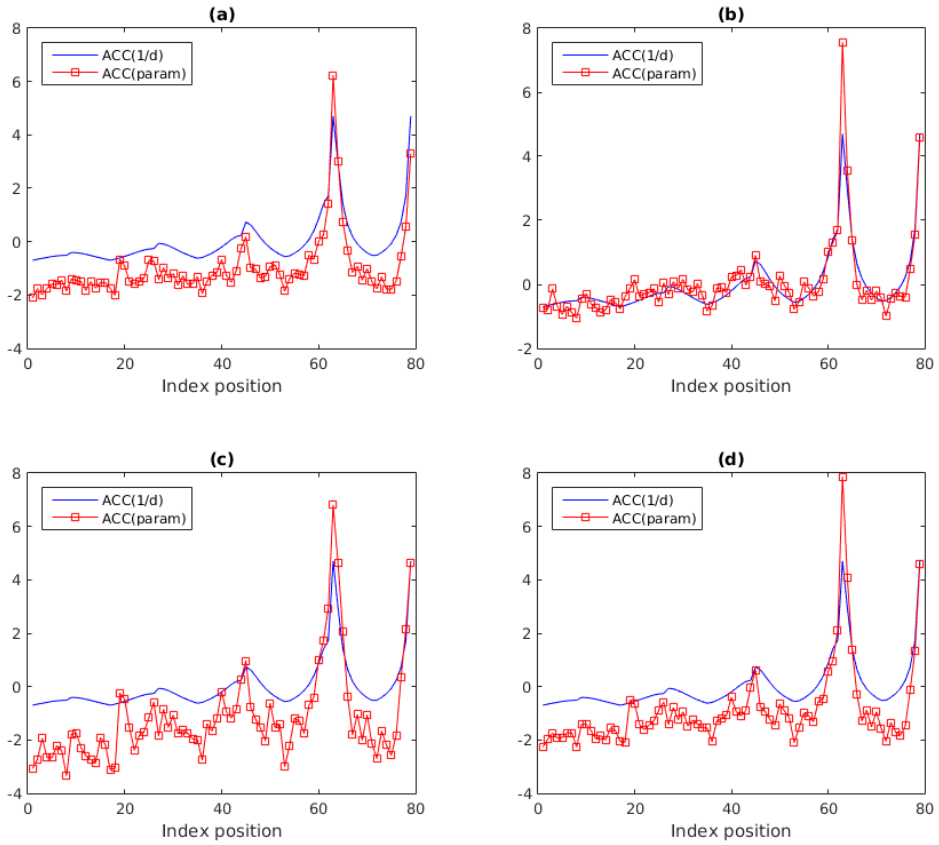



FIGURE 3.9 – Résultat du CCA sur les données de test pour 4 combinaisons de descripteurs : énergie + logV en (a), énergie + SRMR en (b), SRMR + logV en (c) et énergie + logV + SRMR en (d)

TABLE 3.3 – Erreur Quadratique Moyenne entre la référence et les 4 configurations des paramètres

| | **RSB (dB)** | | |
|---|---|---|---|
| **Paramètres** | 13.3 | 6 | -3.4 |
| Énergie + logV | 1.164 | 0.944 | 0.798 |
| Énergie + SRMR | **0.198** | **0.371** | **0.324** |
| SRMR + logV | 2.263 | 0.797 | 0.721 |
| Énergie + SRMR + logV | 1.228 | 1.137 | 1.031 |

#### 3.3.1.2 Généralisation d'un modèle mathématique de l'estimateur de la distance source-microphone

Eu égard aux évaluations des performances des paramètres, l'estimateur de la distance source-microphone est défini de la manière suivante :

$$h_{x1}.\frac{1}{d_i} + h_{x2} \approx h_{y1}.Energie_i + h_{y2}.SRMR_i + h_{y3} \tag{3.16}$$

soit :

$$d_i \approx \frac{h_{x1}}{h_{y1}.Energie_i + h_{y2}.SRMR_i + h_{y3} - h_{x2}} \tag{3.17}$$

Cependant, afin de respecter la contrainte $d \in \mathcal{R}^+$, au calcul direct par la formule 3.17 nous préférerons l'estimation de $d$ par mise en correspondance par la méthode des moindres carrés des valeurs des paramètres avec la vérité-terrain, soit :

$$d_i \approx \mathbf{d}(\underset{j}{\text{argmin}}(h_{x1}.\frac{1}{d_j} + h_{x2} - (h_{y1}.Energie_i + h_{y2}.SRMR_i + h_{y3}))^2) \tag{3.18}$$

Les résultats de l'estimation de la distance, pour le quadrant le plus proche du microphone, sont illustrés sur la figure 3.10, au SNR le plus élevé. Considérant un micro placé à la position (0,0), les positions réelles de la source sonore sont représentées par un point rouge, et l'ensemble possible des positions estimées par un quart de cercle vert centré à la position du microphone et de rayon la distance estimée. Sur les axes des abscisses et des ordonnées sont notées les distances en $x$ et en $y$ au microphone. Les erreurs d'estimations sont quantifiées en terme d'erreur quadratique moyenne et d'erreur moyenne absolue (MAE) sur l'ensemble des 81 positions de la salle, les valeurs de ces indicateurs sont regroupées dans le tableau 3.4.

Table 3.4 – Statistiques sur les erreurs d'estimation de la distance.

| | **RSB (dB)** | | |
|---|---|---|---|
| **Indicateurs** | 13.3 | 6 | -3.4 |
| MSE ($m^2$) | 0.60 | 9.14 | 10.28 |
| MAE ($m$) | 0.88 | 1.16 | 1.31 |

Si l'erreur absolue moyenne augmente peu en fonction du niveau de bruit, l'erreur quadratique moyenne explose dès l'ajout de bruit. La MSE majorant les grosses erreurs, il est possible de conclure que pour des observations bruitées l'estimation reste relativement efficace sur l'ensemble des positions, mais qu'elle peut totalement défaillir, contrairement au contexte non bruité.

Dans cette section, nous avons appris un estimateur de la distance source-microphone à partir d'une combinaison de l'énergie du signal source et du taux de réverbération capté. Nous associons ci-après les observations audio et vidéo qui partagent de possibles positions suffisamment proches au même instant.

## 3.4 Association audio-vidéo par estimation de distance mutuelle

De manière similaire à la fusion par extraction de zones audiovisuelles saillantes, vue en section 3.2.1, nous procédons à l'association des observations audio et vidéo lorsque leurs distances respectives estimées sont suffisamment proches. Soient $y_i^a(t)$ et $y_j^v(t)$ deux observations

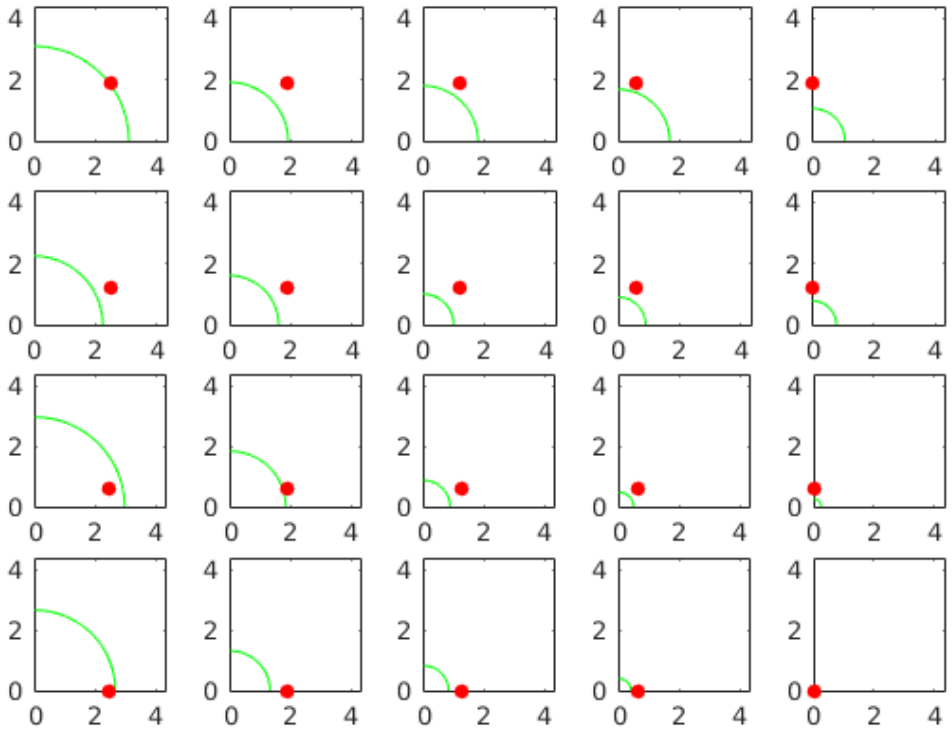


FIGURE 3.10 – Résultat d'estimation de la distance pour 20 positions comprises entre 0 et 4 m.

respectivement audio et vidéo, acquises au temps $t$, $y^{av}(t)$ l'observation audiovisuelle résultant de leur association et $th$ un seuil représentant l'écart toléré entre les distances estimées des deux observations en dessous duquel les observations audio et vidéo peuvent être considérées comme suffisamment proches pour être associées :

$$y^{av}(t) = \begin{cases} \left\{ y_i^a(t), y_j^v(t) \right\}, & \text{si } \left| \mathbf{d}(y_i^a(t)) - \mathbf{d}(y_j^v(t)) \right| \leq th \\ \emptyset, & \text{sinon} \end{cases} \tag{3.19}$$

En faisant varier le seuil $th$ entre 0 et 4m, nous pouvons extraire le taux d'association d'observations audio et vidéo en fonction de la tolérance aux erreurs, soit le nombre de positions où l'association a été réalisée parmi les 81 positions de la salle. En fixant une tolérance d'association de 1m, nous affichons sur la figure 3.11 a), b) et c) les associations (ligne verte) entre les observations vidéo (symbole « plus » rouge) et les observations audio (cercle bleu) aux 3 RSB de notre corpus. L'évolution de cette proportion d'observations fusionnées est illustrée en figure 3.11 d) pour les 3 niveaux de bruits du jeu de données de test. Ainsi pour un RSB de 13,3 dB, en autorisant une erreur d'estimation de 50 cm, la fusion des observations pourra être réalisée sur environ un tiers de la surface de la salle. La valeur de ce seuil est à fixer en fonction du contexte des acquisitions. Dans un lieu peu peuplé, une plus grande tolérance est envisageable, accélérant la fusion des observations.

Nous comparons également avec la méthode présentée dans la section 3.2 en superposant la zone de saillance audiovisuelle, en tirets noirs sur les figures 3.11 a), b) et c). L'intérêt d'une estimation directe de la distance source-microphone est ici bien visible, l'association des observations n'est plus limitée spatialement, donc moins contrainte par le déplacement des usagers.

Nous pourrons ainsi verrouiller les signatures sur des portées plus grandes.

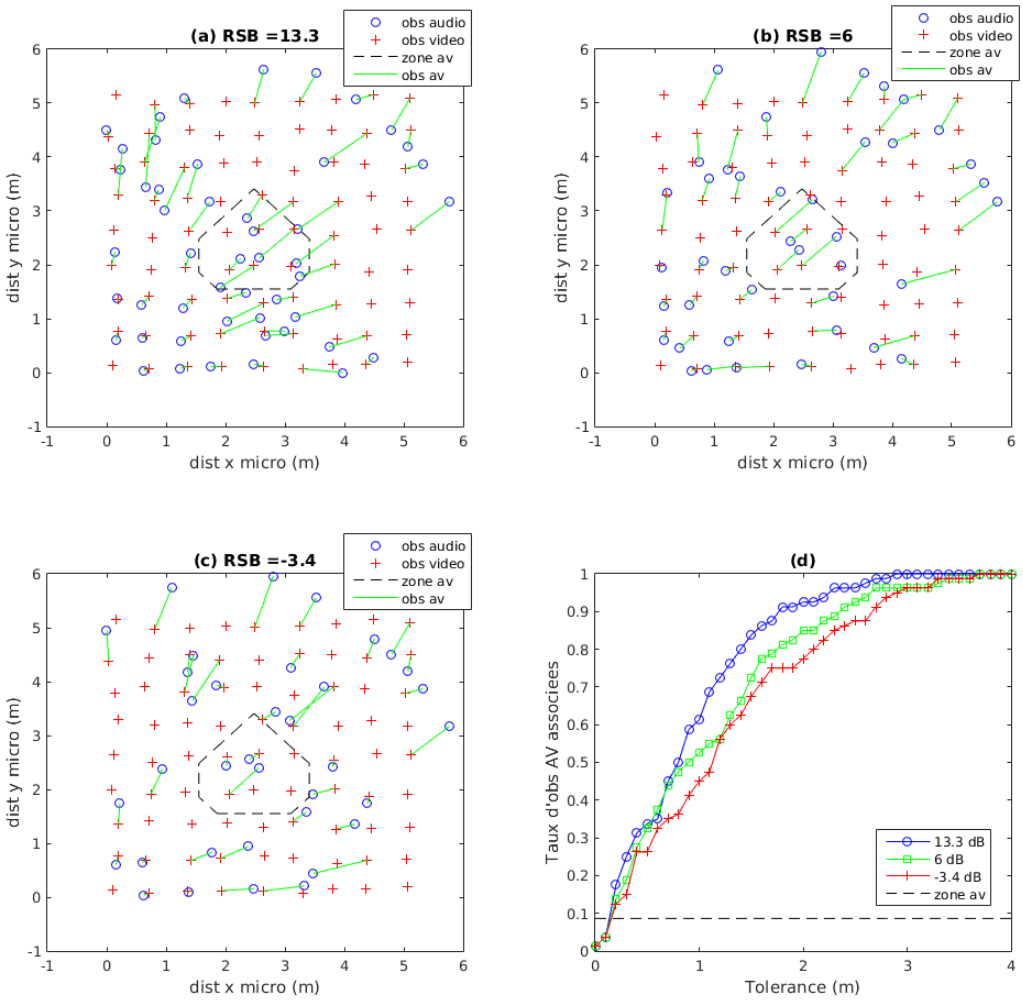



FIGURE 3.11 – Associations des observations audio et vidéo pour 3 niveaux de bruits en a), b) et c) lorsque $th = 1$ et taux d'observations associées en fonction du seuil $th$.

## 3.5 Conclusion

Dans ce chapitre nous avons traité de la fusion des signatures audio et vidéo définies au chapitre précédent, afin d'en créer une signature unique, audiovisuelle. Les deux caractéristiques qui les constituent étant indépendantes, nous les avons associées en fusion tardive, en s'appuyant sur la cohérence spatiale des observations à chaque instant. Si la localisation des observations vidéo est directement réalisable, sous réserve d'une étape de calibration, la localisation de la source sonore nécessiterait un tout autre appareillage que celui à notre disposition. Nous avons alors proposé, évalué et validé deux méthodes pour une estimation de la distance entre une source sonore et un microphone.

Par l'exploitation d'une mesure d'intelligibilité de la parole en environnement réverbéré, le SRMR, nous avons pu dessiner les contours d'une zone de saillance audiovisuelle, autour du microphone, par l'introduction d'un Indice de Proximité Audio Vidéo (IPAV). Ainsi une valeur de l'IPAV supérieure à un seuil assure que les observations audio et vidéo correspondantes se situent dans une zone proximale autour du microphone et qu'elles proviennent de la même personne, permettant la fusion des deux signatures.

Face à la faible portée de ces zones, nous avons exploré des combinaisons de paramètres par l'Analyse Canonique des Corrélations afin d'obtenir une estimation directe de la distance source-microphone et de pouvoir ainsi fusionner les signatures sur tout l'espace expérimental, et non plus sur uniquement certaines zones, augmentant grandement le nombre d'associations d'observations audio-vidéo, et donc de fusion de signatures associées.

Dans le chapitre suivant, nous chercherons à intégrer la cohérence spatio-temporelle des percepts audio et vidéo dans le processus de ré-identification multimodale.

# Chapitre 4

# Apprentissage de signatures audio-visuelles en contexte multi-cibles

**Sommaire**



## Introduction et verrous scientifiques

Dans les chapitres précédents, les tâches relatives à l'apprentissage de signatures audiovisuelles ont été traitées indépendamment à chaque instant $t$ et dans un contexte mono-cible. En effet, hormis la soustraction de l'arrière-plan, tous les outils (détection visuelle, d'activité

vocale, extraction des descripteurs, localisation et mise en correspondance) ne s'appuient que sur la cohérence et la compatibilité spatiale des détections sonores et visuelles. Ce chapitre est une extension portant sur deux niveaux : (i) spatio-temporel, soit en raisonnant sur des observations audio et vidéo acquises à des instants différents, et (ii) multi-cibles, soit par la présence simultanée de plusieurs cibles dans les flux audio et vidéo donc un processus d'association de données plus complexe.

Nous nous intéressons donc ici au problème du suivi multi-cibles audiovisuel. Précédemment, la mise en correspondance des détections sonores et visuelles pour le verrouillage d'une signature audiovisuelle était réalisée sous l'hypothèse de leur forte proximité spatiale, se rapportant ainsi à une classification binaire. Valide dans le cas mono-cible, ou si les cibles sont suffisamment distantes les unes des autres, cette hypothèse se révèle trop lâche lorsque la densité des cibles augmente. L'extension à des scènes encombrées donc multi-cibles pose le problème de l'association des observations ; il est alors opportun de raisonner sur des horizons temporels pour agréger davantage d'information.

Considérant un ensemble d'observations sur une fenêtre temporelle de durée $T$, nous cherchons à associer entre elles les détections appartenant à la même cible, formant ainsi une trajectoire ou « tracklet ». La recherche de la meilleure partition, c'est-à-dire l'ensemble des trajectoires eu égard à l'ensemble des observations, s'appuie sur un modèle dynamique des cibles, ainsi que sur les signatures audio et vidéo inhérentes à chaque cible.

Ce chapitre propose un traqueur multimodal multi-cibles, intégrant les signatures sonores et visuelles vues précédemment dans le but de robustifier leur verrouillage, au sein d'une technique d'association de données de l'état de l'art s'appuyant initialement sur un modèle dynamique des cibles observées. Le chapitre est organisé comme suit : la section 4.1 est consacrée à la problématique du suivi multi-cibles et des approches existantes associées. La section suivante détaille notre approche, son implémentation, et les évaluations sur données synthétiques. Enfin, l'intégration des signatures audiovisuelles des cibles au suivi, ainsi que son évaluation, font l'objet de la section 4.3.

Nous privilégions ici des évaluations sur données simulées pour les raisons suivantes :

— Il n'existe pas de bases de données publiques audio-visuelles compatibles avec nos besoins.
— Nous souhaitons disposer d'une base de données avec vérité terrain.
— Nous souhaitons évaluer la robustesse de notre processus d'apprentissage de signature aux artefacts liés aux observations (non détection des cibles, etc.).

Si les données relatives à la dynamiques sont bien synthétiques, l'intégration des signatures audio et vidéo reposera sur des modèles générés à partir de données réelles, afin d'en faire la meilleure représentation.

## 4.1 Problématique et positionnement de nos travaux

Pour rappel, le suivi consiste à estimer les trajectoires des personnes cibles, en liant les observations correspondantes. La figure 4.1 montre le schéma bloc d'un traqueur visuel multi-cibles dans le plan du sol. Ce chapitre focalise sur la génération des trajectoires, représentée par le bloc rouge, en considérant les blocs intermédiaires comme acquis.

Pour la suite, nous notons :

— $\mathbf{y}_t^j$ : le vecteur contenant la $j$-ème observation, ou mesure, au temps $t$,
— $\mathbf{Y}$ : l'ensemble de toutes les observations sur la durée d'acquisition,
— $\mathbf{x}_t^k$ : le vecteur d'état de l'objet $k$, au temps $t$,
— $\tau_k$ : la $k$-ème trajectoire, contenant la totalité des observations associées à l'objet $k$,

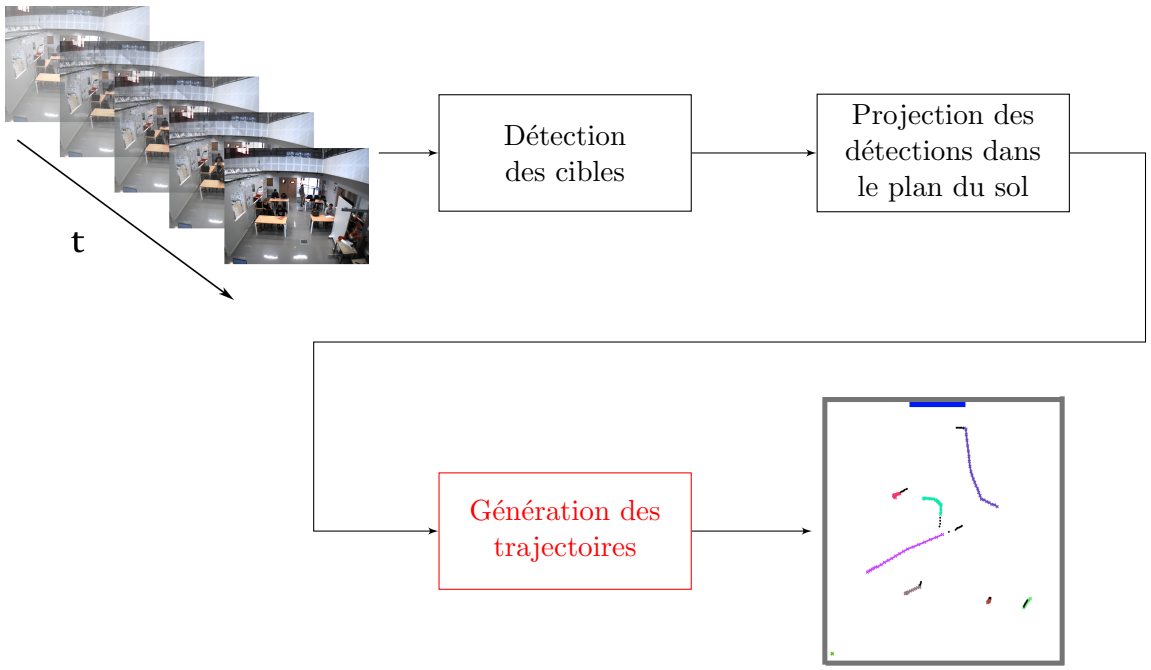



FIGURE 4.1 – Synoptique d'un traqueur visuel multi-cibles depuis un ensemble d'images successives (en haut à gauche) aux trajectoires des cibles inférées dans le plan du sol (en bas à droite).

— $\omega = \{\tau_0, \tau_1, \tau_2, ..., \tau_K\}$ : une partition des observations, regroupant toutes les trajectoires estimée ainsi que l'ensemble des fausses alarmes, noté $\tau_0$,
— $\Omega$ : la collection des différentes partitions estimées.

Pour réaliser la génération des trajectoires, la modalité MOT (pour « *Multiple Object Tracking* ») englobe classiquement :

— une tâche d'association des observations (détections),
— une tâche de filtrage pour estimer l'état propre à chaque cible à l'aide des observations passées et éventuellement présentes,
— une tâche de gestion des étapes de création/terminaison des trajectoires.

### 4.1.1 Etat de l'art et justification de nos choix

Les traqueurs sont catégorisables en terme de (i) nombre de cibles, (ii) fenêtre/horizon temporel(le) considéré(e), ou (iii) stratégies d'association de données. Nous présentons ci-après cette catégorisation.

Nous nous positionnons naturellement dans un cas de suivi multi-cibles (MOT), ainsi le suivi mono-cible (SOT pour « *Single Object Tracking* ») est hors de notre contexte et ne sera pas détaillé ici. Nous retrouverons cependant certaines composantes communes au SOT et au MOT, notamment au niveau du filtrage.

#### 4.1.1.1 Approches en ligne/hors ligne/en logique différée

Le suivi de cible, unique comme multiples, est réalisé à partir d'un ensemble d'observations, ou de mesures, capturées sur une fenêtre temporelle de taille $T$. Les approches existantes diffèrent dans l'utilisation des observations passées, présentes et futures pour l'estimation de l'état des cibles à un instant $t$. Nous catégorisons les approches comme suit (cf. tableau 4.1) :

— Approches en ligne : particulièrement adaptées aux applications où le temps réel est requis, les approches en ligne exploitent uniquement les observations capturées dans l'espace temporel $[0, t]$ pour estimer l'état des cibles en $t$, $t \in [0, T]$. La latence entre acquisition des observations et inférence de l'état est faible ou nulle, mais au détriment de l'absence d'information sur les observations futures, pourtant précieuse, en particulier en suivi multi-cibles.

— Approches hors ligne, ou globales : à l'inverse des approches en ligne, les approches globales vont prendre en considération l'ensemble des observations capturées sur tout l'espace temporel $[0, T]$ pour estimer l'état des cibles à l'instant $t$, $t \in [0, T]$. On infère donc l'état de chaque cible à partir d'un maximum d'information.

— Approches en logique différée, ou *multi-scan* : à l'interface entre les deux, les approches en logique différée adoptent la philosophie du traitement hors ligne sur une fenêtre temporelle glissante : pour estimer l'état des cibles à l'instant $t$, les observations jusqu'à $t + \Delta t$ sont exploitées, $\Delta t$ représentant un futur proche. La latence est ainsi minimisée, tout en s'approchant des performances hors ligne.

Table 4.1 – Spécificités des stratégies SOT/MOT en ligne vs. hors ligne vs. logique différée.

| | **Suivi** | | |
|---|---|---|---|
| **Spécificités** | En ligne | Hors ligne | Logique différée |
| Fenêtre d'observation | $[0, t]$ | $[0, T]$ | $[0, t + \Delta t]$ |
| Méthode | Extension séquentielle des trajectoires avec les observations en $t$ | Association de toutes les observations à $t = T$ | Prolongement des trajectoires avec prise en compte du futur proche |
| Avantages | Adapté au temps réel | Exploitation optimale de l'ensemble d'observations | Compromis entre latence et performance |
| Inconvénients | Ensemble d'observations réduit | Résultats délayés, pas de temps réel | Latence, mais moindre qu'en hors ligne |

#### 4.1.1.2 Approches déterministes/probabilistes

Enfin, le processus d'association de données se catégorise comme suit :

— Approches déterministes : celles-ci cherchent la meilleure observation à associer à chaque trajectoire pour la prolonger, parmi l'ensemble des observations satisfaisant certaines contraintes.

— Approches probabilistes : ces méthodes font intervenir l'estimation des probabilités des hypothèse en remplacement de décisions binaires, permettant ainsi de gérer un grand nombre d'incertitudes.

#### 4.1.1.3 Stratégies usuelles d'association de données (détections) -

Nous présentons par la suite quatre stratégies usuelles d'association de données respectivement en ligne et hors ligne ainsi que respectivement déterministe et probabiliste.

**Global Nearest Neighbors (GNN) -** Cette approche déterministe en ligne s'appuie sur la minimisation d'une fonction de coût dépendante des observations et des états estimés des cibles [FF84]. L'état $x_t^k$ d'une trajectoire $\tau_k$ à l'instant $t$ est prédit par filtrage à l'aide de son état précédent $x_{t-1}^k$ : on propage une densité de probabilité centrée à la position de l'état prédit et dont la matrice de covariance représente la zone d'incertitude de la prédiction.

Par seuillage de cette densité (« *gating* »), on définit ainsi une zone dans laquelle une ou plusieurs observations peuvent être présentes à l'instant $t$. L'approche GNN consiste alors à assigner un coût à chaque observation, communément basé sur la distance (euclidienne, Mahalanobis) entre la prédiction et l'observation, mais pouvant également être enrichi, notamment par la similarité des apparences.

L'association des données aux trajectoires est ensuite réalisée par minimisation de ces fonctions de coût, sous la contrainte d'unicité d'affectation d'une observation à une trajectoire. Ce problème d'assignation peut être résolu par méthode gloutonne ou en utilisation des méthodes d'optimisation comme l'algorithme hongrois [Mun57].

**Joint Probabilistic Association (JPDA) -** Cette méthode réalise également le suivi multi-cibles en ligne mais de manière probabiliste [FBSS83]. Elle reprend une démarche similaire au GNN de prédiction puis gating, mais diffère dans le choix des observations. Là où le GNN n'en choisit qu'une et l'assigne à la trajectoire, l'ensemble des observations dans la zone définie par le gating contribuent au prolongement de la trajectoire.

Pour chaque observation de la zone, on calcule la probabilité d'appariement aux trajectoires, dans toutes les partitions possibles, sur les mêmes indices que la fonction de coût du GNN. Les trajectoires sont ensuite prolongées par filtrage de Kalman à observations multiples, où on utilise une somme des résidus pondérées par ces probabilités.

**Multiple Hypothesis Tracker (MHT) -** Contrairement au GNN et au JPDA, le MHT traite le suivi multi-cibles hors ligne, ou en logique différée [Rei79]. Dans cette approche déterministe, toutes les partitions possibles des trajectoires sont testées, sous certaines hypothèse de voisinage d'observations successives.

À partir de l'observation initiale, cette collection de partitions se structure alors en arbre de possibilités. Afin d'en réduire la taille, une opération d'élagage (« *pruning* ») est réalisée. La probabilité de chaque partition est calculée comme la somme des probabilités des trajectoires la constituant, celles-ci étant calculées à partir d'une vraisemblance de la dynamique par filtrage.

L'association des observations aux trajectoires est alors réalisée en choisissant la partition la plus probable parmi la collection.

**Markov Chain Monte Carlo for Data Association (MCMCDA) -** Cette approche, initiée dans [Oh+08], traite également le problème de suivi multi-cibles hors ligne ou en logique différée mais en utilisant une approche probabiliste. Elle utilise un processus MCMC pour générer de manière itérative des partitions aléatoires successives sur toute la durée d'observation.

La probabilité de la partition générée à une itération $i$ est calculée, en se basant sur le modèle dynamique, et comparée à la probabilité de la partition à l'itération $i-1$. On décide d'accepter cette partition ou de la rejeter en fonction de ce rapport des probabilités, puis la

nouvelle partition (celle à l'itération $i$ en cas d'acceptation, et celle à l'itération $i-1$ en cas de rejet) est remise en cause à l'itération $i+1$. Selon certaines conditions, on peut alors montrer la convergence vers la partition optimale.

**Synthèse** Eu égard à notre contexte applicatif, soit un suivi multi-cibles non contraint au temps réel, mais toutefois à délai court pour la prise de décision, nous privilégions un traqueur MOT en logique différée pour le compromis latence/performance.

Les quatre stratégies d'association de données présentées réalisent le suivi de manière déterministe en ligne (GNN), probabiliste en ligne (JPDA), déterministe hors ligne (MHT) et probabiliste hors ligne (MCMCDA). Le GNN et le JPDA souffrent non seulement de l'absence de remise en cause des partitions, dû au caractère en ligne des approches, mais nécessitent une supervision pour la gestion de la création et de la terminaison des trajectoire. Dans notre contexte applicatif, le nombre et l'apparition des cibles sont inconnus, ces méthodes sont alors inadaptées à notre application.

L'approche MHT gère ces inconnues et offre la possibilité d'exploiter l'information de toute la durée d'observation par son caractère hors-ligne. Le coût machine de cette méthode est cependant très élevé et subit une croissance exponentielle avec le nombre de cibles à suivre. L'opération de pruning permet de réduire ce coût mais au détriment du risque de suppression de la bonne partition. Nous sélectionnons alors l'approche MCMCDA, en accord avec le choix de l'emploi d'une méthode probabiliste hors ligne/en logique différée, qui combine les forces du MHT et une gestion moins coûteuse des grands nombres de cibles. Cette technique sera détaillée en §4.2.

L'état des cibles sera classiquement géré par un filtrage de Kalman, donc également un formalisme probabiliste. La section suivante en rappelle les grandes lignes.

### 4.1.2 Rappels sur le filtrage de Kalman

Le principe est d'estimer une suite d'états $\mathbf{x} = \{\mathbf{x}_t\}_{t=1:T}$ dans lesquels évolue la cible au cours de la période d'observation $[0, T]$. Les états étant cachés, à savoir non mesurables directement, ils seront estimés à partir d'un ensemble d'observations $\mathbf{y} = \{\mathbf{y}_t\}_{t=1:T}$. L'évolution du système peut être décrite à l'aide de deux modèles : un modèle d'état et un modèle d'observation. Ils sont explicités par la suite dans le cas où cette évolution est supposée linéaire.

**Modèle d'état -** Il représente la dynamique du système, l'évolution théorique de l'état par rapport au passé. Un modèle fréquemment utilisé en suivi est un modèle markovien d'ordre 1 (l'état de la cible en $t+1$ ne dépend que de l'état en $t$, et non des états antérieurs, en $[0, t-1]$) :

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t) + w_t, \qquad \text{pour } t = t_i, \ldots, t_{f-1} \tag{4.1}$$

en considérant :

- $[t_i, t_f] \subset [1, T]$ la période pendant laquelle l'objet évolue dans la région d'observation,
- $F : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ la dynamique, à temps discret de l'objet avec $n_x$ la dimension de la variable d'état $\mathbf{x}$,
- $w_t \in \mathbb{R}^{n_x}$ un processus aléatoire modélisant un bruit de prédiction.

**Modèle d'observation -** Il exprime le comportement des mesures par rapport aux états. Chaque mesure $\mathbf{y}_t^i$ ne dépend que de l'état $\mathbf{x}_t$, et les mesures sont indépendantes entre elles. Une mesure pouvant être une réelle observation de l'objet, comme une fausse alarme, nous exprimons

son modèle comme suit :

$$\mathbf{y}_t^j = \begin{cases} H(\mathbf{x}_t) + v_t & \text{si la j-ème observation est une mesure de } \mathbf{x}_t \\ u_t & \text{sinon} \end{cases} \tag{4.2}$$

avec :

- $n_t$ le nombre d'observations au temps $t$, aussi bien les vraies observations bruitées que les fausses alarmes,
- $H : \mathbb{R}^{n_y} \to \mathbb{R}^{n_y}$ la fonction d'observation, où $n_y$ est la dimension des vecteurs d'observation,
- $y_t^j \in \mathbb{R}^{n_y}$ la $j$-ème observation au temps $t$ pour $j = 1, \ldots, n_t$.

**Filtre de Kalman -** Nous présentons ici succinctement le fonctionnement d'un filtre de Kalman, car il est utilisé dans toutes les approches présentée ensuite, et particulièrement dans celle que nous retenons. Introduit en [Kal60], il est une approximation linéaire du filtrage bayésien. Il estime récursivement l'état du système à travers une étape de prédiction ($\mathbf{x}_t$ prédit à partir de $\mathbf{x}_{t-1}$) et de mise à jour ($\mathbf{x}_t$ corrigé par $\mathbf{y}_t$). Ces étapes s'appuient sur l'estimation de densités de probabilités, eu égard aux modèles d'état et d'observation ci-dessus, centrées sur les états, respectivement prédits et mis à jour, et dont les matrices de covariance, respectivement d'erreur de prédiction, et d'erreur de mise à jour, sont estimées récursivement.

L'inconvénient principal de ce filtre est sa limitation aux systèmes linéaires. Il a été étendu aux systèmes non-linéaires dans [SJJ97], mais dans notre contexte néanmoins, nous pouvons accepter cette hypothèse de linéarité et utiliser un filtre de Kalman classique.

### 4.1.3 Métriques d'évaluation

Les métriques CLEARMOT [BS08] sont couramment usitées pour évaluer les traqueurs MOT. Elles dépendent du nombre de Faux Négatifs (FN, le nombre d'associations manquées), du nombre de Faux Positifs (FP, le nombre de fausses associations réalisées), du nombre de mauvaises associations (IDS pour ID Switch, le nombre d'observations appartenant à une trajectoire et associées à une autre) ainsi qu'un score de précision (dépendant de la distance entre les trajectoires estimées $\tau_k$ et la vérité terrain $\tau_k^*$). Ces métriques se formulent ainsi :

- **Multiple Object Tracking Accuracy (MOTA) :** l'exactitude du suivi représente le nombre de bonnes associations sur le nombre total $P_{\mathcal{O}}$ de positions des cibles de la vérité terrain :
$$MOTA = 1 - \frac{FN + FP + IDS}{P_{\mathcal{O}}} \tag{4.3}$$
- **Multiple Object Tracking Precision (MOTP) :** la précision du suivi représente la distance moyenne des trajectoires estimées $\{\tau_k\}_{k=1,\ldots,K}$ à la vérité terrain $\{\tau_k^*\}_{k=1,\ldots,K}$ :
$$MOTP = \frac{\sum_k \sum_n ||\tau_k(n), \tau_k^*(n)||}{\sum_k |\tau_k|} \tag{4.4}$$

avec $||.,.||$ la distance euclidienne, et $|.|$ le cardinal, soit le nombre d'éléments dans la trajectoire. Les performances d'un traqueur sont évaluées via ces deux critères, voire seulement le MOTA si nous évaluons uniquement l'association des détections entre elles (sans estimation de précision de localisation).

## 4.2 MCMCDA : association de données par MCMC

Eu égard à nos propos préalables, notre choix s'est porté sur le MCMCDA pour ses performances en environnement dense, son traitement en logique différée et son maintien du nombre inconnu de cibles ainsi que de leur identité. Cette section rappelle son formalisme, notre implémentation puis validation sur données synthétiques.

La technique MCMCDA réalise l'association des observations en employant une méthode de Monte-Carlo par chaînes de Markov (MCMC). Une partition des observations $\omega$ (regroupement en trajectoires et fausses alarmes) est tirée aléatoirement à travers une distribution de mouvements possibles des trajectoires et la partition tirée à l'itération précédente. La vraisemblance de la nouvelle partition est ensuite comparée à celle de la précédente, et est soit acceptée soit refusée. Le processus est répété sur un nombre $n_{mc}$ d'itérations et converge ainsi vers la partition optimale.

### 4.2.1 Formalisation du MCMC

Résumons ici le principe général de cette technique.

#### 4.2.1.1 Concepts généraux et intérêts

Un calcul de probabilité nécessite généralement l'intégration d'une distribution qui, si elle est réalisable, se révèle généralement fastidieuse. En effet, en reprenant le théorème de Bayes, si nous voulons exprimer la probabilité *a posteriori* d'avoir la partition *omega* connaissant l'ensemble d'observation $\mathbf{Y}$, $P(\omega|\mathbf{Y})$, nous avons :

$$P(\omega|\mathbf{Y}) = \frac{P(\mathbf{Y}|\omega)P(\omega)}{P(\mathbf{Y})} \tag{4.5}$$

Il est possible de calculer directement les éléments du numérateur, la fonction de vraisemblance $P(\mathbf{Y}|\omega)$ et la loi marginale $P(\omega)$, mais le calcul du dénominateur nécessite l'intégration suivante :

$$P(\mathbf{Y}) = \int_{\Omega} P(\mathbf{Y}|\omega')P(\omega')d\omega' \tag{4.6}$$

L'intérêt des méthodes MCMC est de s'affranchir de ce calcul, et de tirer des échantillons successifs $(\omega_1, \omega_2, \ldots, \omega_n)$ de la distribution, tels qu'ils forment une chaîne de Markov (la génération de $\omega_i$ ne dépend que de $\omega_{i-1}$, et non des $\omega_{0,..,i-2}$) qui converge vers la distribution stationnaire $P(\omega|\mathbf{Y})$. Selon le théorème ergodique la chaîne doit alors être irréductible, récurrente positive et apériodique.

#### 4.2.1.2 Algorithme de Metropolis-Hastings

Dans cette classe de méthodes d'échantillonnage, nous pouvons relever deux méthodes notables : l'algorithme de Metropolis-Hastings et l'échantillonneur de Gibbs. Introduit en [Met+53] et généralisé en [Has70], l'algorithme de Metropolis-Hastings est un processus itératif en deux étapes. On échantillonne une distribution $q(\omega', \omega)$ de probabilité de la partition $\omega'$ par rapport à la partition courante $\omega$. La décision sur l'échantillon est ensuite prise par un critère

d'acceptation-rejet $A(\omega', \omega)$, en étudiant le rapport des probabilités des deux états $\omega$ et $\omega'$ de la chaîne :

$$A(\omega', \omega) = min\left(1, \frac{P(\omega'|\mathbf{Y})q(\omega', \omega)}{P(\omega|\mathbf{Y})q(\omega, \omega')}\right) \tag{4.7}$$

Les étapes de la méthode sont présentées dans l'algorithme 1.

**Algorithm 1:** Metropolis-Hastings

**Input :** $Y$, $n_{mc}$, $\omega_{init}$
**Output:** $\hat{\omega} = \text{argmax}_n(P(\omega(n)|Y))$

$\omega \leftarrow \omega_{init}$
**for** $n = 1$ **to** $n_{mc}$ **do**
  proposer $\omega'$ selon $\omega$ par $q$
  tirer $U \sim \text{Unif}[0, 1]$
  **if** $U < A(\omega, \omega')$ **then**
    $\omega \leftarrow \omega'$
  **end**
  $\omega(n) \leftarrow \omega$
**end**

L'échantillonneur de Gibbs [GG84] propose lui de diviser l'échantillonnage d'une variable $\omega$ de grande dimension $N$ et de tirer séparément des échantillons à une seule dimension $[\omega^1, \omega^2, ..., \omega^N]$, permettant ainsi de traiter simplement des problèmes en grande dimension. Nous ne le détaillerons pas ici, car seul l'algorithme de Metropolis-Hastings est employé dans le MCMCDA.

Nous présentons ensuite l'adaptation de cet algorithme au problème d'association de données.

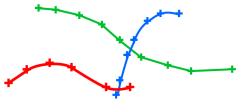
### 4.2.2 Formalisation adaptée au MOT

L'approche MCMCDA propose un traitement global de l'ensemble $\mathbf{Y}$ des observations détectées dans la région considérée et au cours de toute la période d'observation $[0, T]$. Eu égard à l'agencement spatio-temporel de $\mathbf{Y}$, la variable échantillonnée $\omega$ est la partition des observations, soit leur répartition au sein de trajectoires $\{\tau_i\}_{i=1,...,K}$. Les observations non affectées à une trajectoire sont regroupées dans l'ensemble des fausses alarmes $\tau_0$. La partition doit respecter les contraintes suivantes :

— unicité d'affectation des observations : $\tau_i \cap \tau_j = \emptyset$, pour $i \neq j$
— étiquetage de toutes les observations : $\cup_{k=0}^{K} \tau_k = \mathbf{Y}$
— au maximum une observation par instant et par trajectoire : $|\tau_k \cap \mathbf{Y}(t)| \leq 1$
— longueur minimum d'une trajectoire : $|\tau_k| \geq 2$ pour $k = 1, \ldots, K$

Avec ces contraintes, définissons la distribution $q(\omega', \omega)$, dont les échantillons forment la chaîne de Markov convergeant vers la partition optimale des observations. Nous synthétisons dans le tableau 4.2 l'ensemble des éléments traités en MOT ainsi que les notations usitées.

TABLE 4.2 – Notations et illustrations des éléments traités en suivi multi-cibles.

| Élément | Notation | Illustration |
|---|---|---|
| Ensemble des observations | $\mathbf{Y}$ |  |
| $n$-ème observation de la $k$-ème trajectoire | $\mathbf{y}_n^k$ ou $\tau_k(n)$ |  |
| $k$-ème trajectoire | $\tau_k$ |  |
| Partition | $\omega$ |  |

#### 4.2.2.1 Mouvements sur les trajectoires

Proposer une partition $\omega'$, à partir d'une partition existante $\omega$, consiste à effectuer des opérations sur les trajectoires sous-jacentes. Afin de respecter les propriétés d'irréductibilité, de récurrence positive et d'apériodicité, 5 types d'opérations sont proposés : création/suppression, division/fusion, extension/réduction, mise à jour et inversion. Ces opérations sont illustrées sur la figure 4.2 et décrites par la suite.

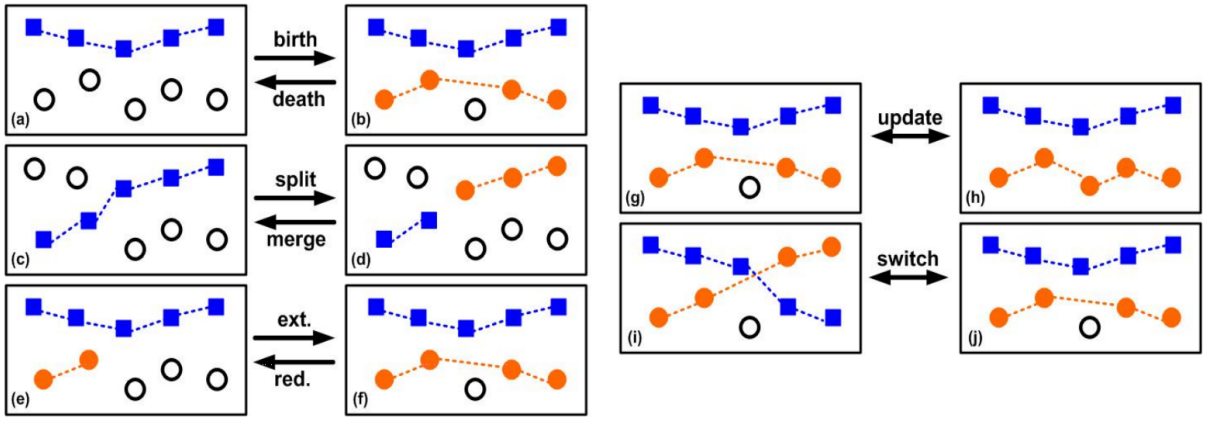



FIGURE 4.2 – Illustrations des mouvements de trajectoires. Les lignes et les formes géométriques colorées représentent respectivement les trajectoires et leurs observations. Les cercles noirs représentent les fausses alarmes. Figure extraite de [Oh+08].

**Création/Suppression -** La création d'une trajectoire est effectuée en associant au minimum deux observations (préalablement étiquetées comme de fausses alarmes) qui respectent

une contrainte de voisinage spatio-temporel. Soient $d_{max}$ la durée maximale autorisée entre deux observations (*e.g.* pour $d_{max}$=4, nous autorisons 3 non-détections successives) et $v_{max}$ la vitesse maximale autorisée des cibles. La création d'une trajectoire s'apparente alors à une marche aléatoire sur les observations se trouvant dans le voisinage $L_d$, initiée à un temps $t_i$ tiré uniformément sur [0,T-1] : on tire aléatoirement une observation dans le voisinage spatio-temporel de l'observation courante, puis on se place sur l'observation choisie et on répète le processus. Le voisinage de la $j$-ème observation à l'instant $t$, $y_t^j$, est défini comme suit :

$$L_d(y_t^j) = \{y_{t+d}^k \quad | \quad ||y_{t+d}^k - y_t^j|| \leq d.v_{max}\} \tag{4.8}$$

L'ajout de nouvelles observations, tirées aléatoirement dans $L_d$, à la trajectoire se poursuit tant que le voisinage contient au moins une observation candidate. À chaque nouvelle association, la trajectoire a une probabilité de terminaison $\gamma < 1$. On tire un échantillon $z$ d'une loi uniforme sur $[0, 1]$ et on interrompt prématurément la marche aléatoire si $z < \gamma$. La nouvelle trajectoire $\tau_k$ est ajoutée à la partition $\omega$ et les observations associées à la trajectoire sont retirées de l'ensemble $\tau_0$ contenant les fausses alarmes. La suppression d'une trajectoire effectue l'opération inverse : une trajectoire est tirée aléatoirement dans la partition, en est effacée, et les observations qui la composaient sont transférées dans $\tau_0$.

**Division/fusion -** La division d'une trajectoire est réalisée selon les étapes suivantes : une trajectoire $\tau_k$ est sélectionnée aléatoirement dans la partition $\omega$ telle que $|\tau_k| \geq 4$, puis un instant $t$ est tiré tel que $|\tau_k(t_i, t-1)| \geq 2$ et $|\tau_k(t, t_f)| \geq 2$ avec $t_i$ et $t_f$ les temps respectivement de début et de fin de $\tau_k$. La trajectoire d'origine $\tau_k$ est ensuite retirée de $\omega$ et les trajectoires $\tau_{k_1} = \tau_k(t_i, \ldots, t-1)$ et $\tau_{k_2} = \tau_k(t, \ldots, t_f)$ y sont ajoutées.

L'opération inverse de la division, la fusion, requiert deux trajectoires $\tau_{k_1}$ et $\tau_{k_2}$ dont les extrémités (fin pour $\tau_{k_1}$ et début pour $\tau_{k_2}$) sont dans le même voisinage : $\tau_{k_1}^{t_f} \in L_d(\tau_{k_2}^{t_i})$ et $\tau_{k_2}^{t_i} \in L_d(\tau_{k_1}^{t_f})$. Un couple $(\tau_{k_1}, \tau_{k_2})$ de trajectoires respectant cette contrainte de voisinage est tiré aléatoirement parmi tous les couples candidats et les deux trajectoires sont supprimées de la partition $\omega$. Enfin nous incluons à cette dernière la nouvelle trajectoire $\tau_k = \tau_{k_1} \cap \tau_{k_2}$.

**Extension/Réduction -** Pour réaliser une extension, nous reproduisons la démarche de création d'une trajectoire depuis une extrémité d'une trajectoire existante, choisie aléatoirement dans la partition. La trajectoire peut être étendue en partant de sa dernière observation, ou de sa première. Dans ce dernier cas, la marche aléatoire est inversée en parcourant le temps de manière décroissante.

La réduction d'une trajectoire $\tau_k$ consiste à réassigner successivement les observations qui la composent à l'ensemble des fausses alarmes $\tau_0$. À chaque réaffectation, et tant que $|\tau_k| \geq 2$, nous testons l'hypothèse d'interrompre la réduction, comme pour les mouvements précédents.

**Mise à jour -** Nous appelons mise à jour l'insertion d'une observation $y_t^0$ de $\tau_0$ dans une trajectoire $\tau_k$ et le retrait de l'observation $y_t^k \in \tau_k$. Cette opération est réalisable à condition que $y_t^0 \in L_d(y_t^k)$. Comme pour les autres mouvements, une trajectoire est tirée aléatoirement de la partition $\omega$, puis une de ses observations respectant la contrainte de voisinage est sélectionnée.

**Interversion -** Cette dernière opération vise à échanger l'identité de deux trajectoires $\tau_{k_1}$ et $\tau_{k_2}$ à leur croisement. Elle nécessite que l'observation d'une première trajectoire $\tau_{k_1}$ à un instant $t$ soit dans le voisinage de l'observation à $t+1$ d'une autre trajectoire $\tau_{k_2}$, et inversement :

$y_t^{k_1} \in L_d(y_{t+1}^{k_2})$ et $y_t^{k_2} \in L_d(y_{t+1}^{k_1})$. Les observations de $\tau_{k_1}$ ultérieures à $t$ sont alors affectées à $\tau_{k_2}$ et les observations de $\tau k_2$ ultérieures à $t$ sont affectées à $\tau_{k_1}$. Les nouvelles versions des trajectoires sont alors :

$$\tau_{k_1} = \{\tau_{k_1}(t_i^{k_1}), ..., \tau_{k_1}(t), \tau_{k_2}(t+1), ..., \tau_{k_2}(t_f^{k_2})\} \tag{4.9}$$

$$\tau_{k_2} = \{\tau_{k_2}(t_i^{k_2}), ..., \tau_{k_2}(t), \tau_{k_1}(t+1), ..., \tau_{k_1}(t_f^{k_1})\} \tag{4.10}$$

L'ensemble des opérations, leurs index et l'état des trajectoires avant/après application, est résumé dans le tableau 4.3.

TABLE 4.3 – Opérations sur les trajectoires

| **Opération** | **Index** | **Entrée** | **Sortie** |
|---|---|---|---|
| Création | 1 | $\emptyset$ | $\tau_k$ |
| Suppression | 2 | $\tau_k$ | $\emptyset$ |
| Extension | 3 | $\tau_k$ | $\tau_k$ |
| Réduction | 4 | $\tau_k$ | $\tau_k$ |
| Mise à jour | 5 | $\tau_k$ | $\tau_k$ |
| Division | 6 | $\tau_k$ | $\tau_{k_1}$, $\tau_{k_2}$ |
| Fusion | 7 | $\tau_{k_1}$, $\tau_{k_2}$ | $\tau_k$ |
| Interversion | 8 | $\tau_{k_1}$, $\tau_{k_2}$ | $\tau_{k_2}$, $\tau_{k_1}$ |

En pratique, pour effectuer la proposition $q(\omega', \omega)$, l'index d'un des 8 mouvements possibles est choisi aléatoirement à l'aide d'une distribution $\xi_K(\omega)$ définie comme suit :

$$\xi_K(\omega) : \begin{cases} = 1 & \text{si} \quad |\omega| = 0 \\ \sim \mathcal{U}(1,6) & \text{si} \quad |\omega| = 1 \\ \sim \mathcal{U}(1,8) & \text{si} \quad |\omega| \geq 1 \end{cases} \tag{4.11}$$

avec $\mathcal{U}(a,b)$ la distribution uniforme des entiers entre $a$ et $b$.

#### 4.2.2.2 Vraisemblance de la partition

Pour accepter ou refuser la nouvelle partition $\omega'$ créée à partir de la partition courante $\omega$ subissant une opération $q(\omega', \omega)$, il est nécessaire de pouvoir estimer sa vraisemblance aux observations $P(\omega'|\mathbf{Y})$ relative à la vraisemblance de la partition courante $P(\omega|\mathbf{Y})$. L'intérêt de l'approche MCMC est de n'avoir à les calculer qu'à une constante multiplicative près :

$$P(\omega|\mathbf{Y}) \propto P(\mathbf{Y}|\omega) \prod_{t=1}^{T} p_z^{z_t}(1-p_z)^{c_t} p_d^{d_t}(1-p_d)^{g_t} \lambda_b^{a_t} \lambda_f^{f_t} \tag{4.12}$$

avec $p_z$ la probabilité de terminaison d'une trajectoire, $z_t$ le nombre de trajectoires se terminant à l'instant $t$, $c_t$ le nombre de trajectoires se poursuivant à $t$, $p_d$ la probabilité de détection des observations, $d_t$ le nombre d'observations détectées à l'instant $t$, $g_t$ le nombre d'observations non détectées à $t$, $\lambda_b$ le taux de création de trajectoires à chaque instant, $a_t$ le nombre de trajectoires apparues à $t$, $\lambda_f$ le taux de fausses alarmes à chaque instant et $f_t$ le nombre de fausses alarmes à l'instant $t$.

Le terme $P(\mathbf{Y}|\omega)$ représente quant à lui la vraisemblance de l'ensemble d'observations eu égard à la partition proposée. En reprenant les équations du filtre de Kalman, l'écart entre

les observations et les états estimés est exprimé à travers la matrice de covariance d'erreur de prédiction. Pour chaque observation $y_t^k$, dans une trajectoire $\tau_k$ et $x_t^k$ l'état correspondant estimé, $P(y_t^k|x_t^k)$ suit une loi normale bidimensionnelle centrée en $x_t^k$, et dont la matrice de covariance correspond à une fonction de l'erreur de prédiction. La vraisemblance totale $P(\mathbf{Y}|\omega)$ est le produit de toutes les vraisemblances des observations :

$$P(\mathbf{Y}|\omega) = \prod_{\tau=\tau_1}^{|\omega|} \prod_{t=1}^{|\tau|} \mathcal{N}(\tau(t)|\mathbf{x}_t(\tau), B_t(\tau)) \tag{4.13}$$

où $B_t(\tau)$ est la matrice de covariance de la $t$-ème observation de la trajectoire $\tau$, estimée à l'aide de la matrice d'erreur de prédiction du filtre de Kalman.

**Implémentation -** Nous avons entièrement implémenté cet algorithme en C++ à l'aide des bibliothèques OpenCV [13] pour le filtrage des observations et Boost [14] pour la gestion de l'échantillonnage des différentes distributions. Par la suite, nous évaluons ses performances sur plusieurs scenarii à partir de données synthétiques.

### 4.2.3 Évaluations sur données simulées

L'efficacité d'une méthode d'association de données dépend de sa capacité à associer deux observations dérivées de la même trajectoire, à dissocier deux observations issues de trajectoires différentes (ou de fausses alarmes) et à conserver l'identité des trajectoires, soit limiter le nombre de fragmentations. Ces objectifs sont d'autant plus contraignants dans les cas suivants :

— une grande densité de trajectoires,
— un taux élevé de fausses alarmes,
— une faible probabilité de détection des observations.

Nous évaluons alors les performances du MCMCDA sur des données synthétiques modélisant ces cas critiques, afin de valider notre implémentation et d'en caractériser fonctionnement aux limites.

**Critères utilisés -** Nous générons trois scenarii expérimentaux dédiés aux trois situations pré-citées. Notre démarche s'inspire de [Oh+08]. Deux critères supplémentaires sont considérés :

— Normalized Correct Associations (NCA) : le rapport entre le nombre d'associations correctes (CA) dans la partition proposée $\omega$ et le nombre total d'associations (SA) dans la vraie partition $\omega^*$. Ce critère est à maximiser.

$$NCA(\omega) = \frac{|CA(\omega)|}{|SA(\omega^*)|} \tag{4.14}$$

— Incorrect to Correct Association Ratio (ICAR) : le rapport entre le nombre d'erreurs d'association et le nombre d'associations correctes. Ce critère est à minimiser.

$$ICAR(\omega) = \frac{|SA(\omega)| - |CA(\omega)|}{|CA(\omega)|} \tag{4.15}$$

13. OpenCV : `http://opencv.org/`
14. Boost : `http://www.boost.org/`

Ces critères sont assimilables au critère de rappel pour le critère NCA, et comme un pseudo inverse d'un critère de précision pour le critère ICAR (l'inverse formel de la précision aurait opposé l'ensemble des associations aux associations correctes, là où le critère ICAR n'y oppose que les associations incorrectes). Notre privilégions ici la confiance des associations, plutôt que leur quantité, et focalisons donc sur la minimisation du critère ICAR.

#### 4.2.3.1 Scénario 1 : variation du nombre de trajectoires

Cette première expérience vise à évaluer les performances de l'association de données en environnement de plus en plus dense. Considérons une région d'observation $R$ de taille $[1000 \times 1000]$ dans laquelle $K$ cibles évoluent au cours de la période d'observation $[0, T]$, $T = 30$ unités de temps. Les mesures associées à chaque trajectoire sont toutes détectées, et des fausses alarmes sont ajoutées à l'ensemble des observations, uniformément réparties sur $R$ et avec une probabilité d'apparition par volume $\lambda_f V = 1$. Les vitesses maximales des cibles sont fixées à $v_{max} = 50$ unités de distance par unité de temps. Les temps et positions d'apparition des trajectoires sont tirés aléatoirement sur $[0, T-1]$, une trajectoire devant toujours contenir au moins deux observations. Enfin, elles se terminent avec une probabilité de disparition $\gamma$. Pour chaque test nous utilisons 10000 itérations.

Dans ce scénario, nous faisons varier le nombre de trajectoires en conservant les autres paramètres fixes, augmentant ainsi la densité des observations sur la région $R$. L'association de données est réalisée pour $K = [5, 10, 20, 30, 40, 50, 75, 100]$. Des exemples de trajectoires ainsi produites sont illustrés en figure 4.3 (toutes les observations de $[0, T]$ sont ici agrégées).

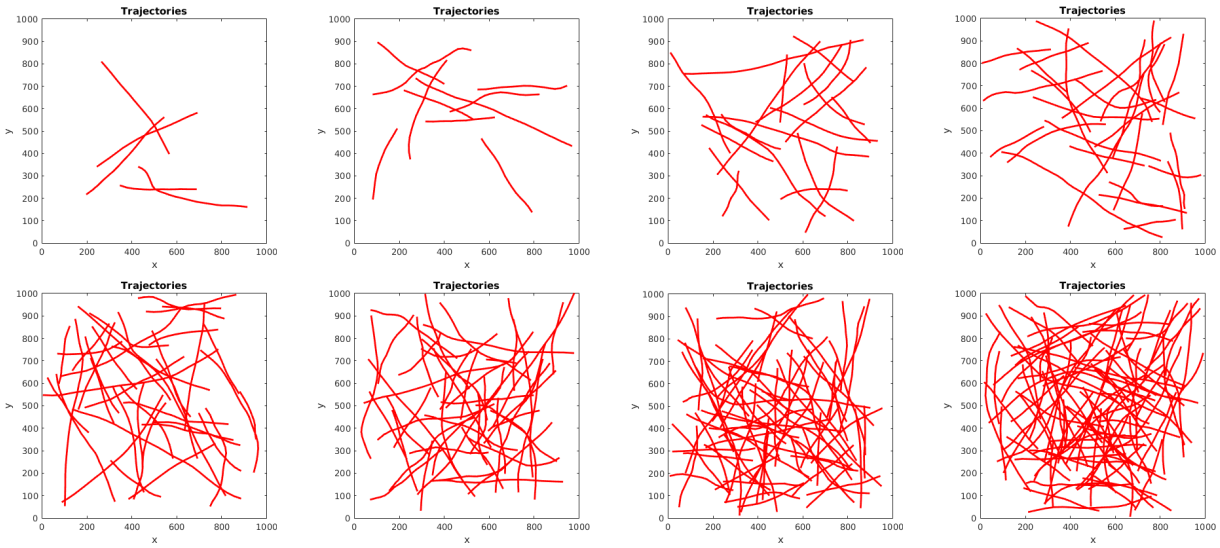



FIGURE 4.3 – De haut-gauche à bas-droit : scenarii de test comportant un nombre variable de $K$ trajectoires, $K = [5, 10, 20, 30, 40, 50, 75, 100]$.

**Analyse qualitative des erreurs -** Les erreurs possibles en association de données sont la non association de deux observations appartenant à la même trajectoire (faux négatif) et l'association de deux observations appartenant à deux trajectoires différentes ou étant des fausses alarmes (faux positif). Ces erreurs peuvent alors causer l'interruption prématurée de trajectoires, mais également les fragmenter (deux trajectoires ou plus estimées au lieu d'une unique), ne

pas associer ponctuellement une bonne observation ou encore intervertir deux trajectoires. Des exemples de fragmentation, d'observation manquée et de ID Switch sont illustrés figure 4.4.

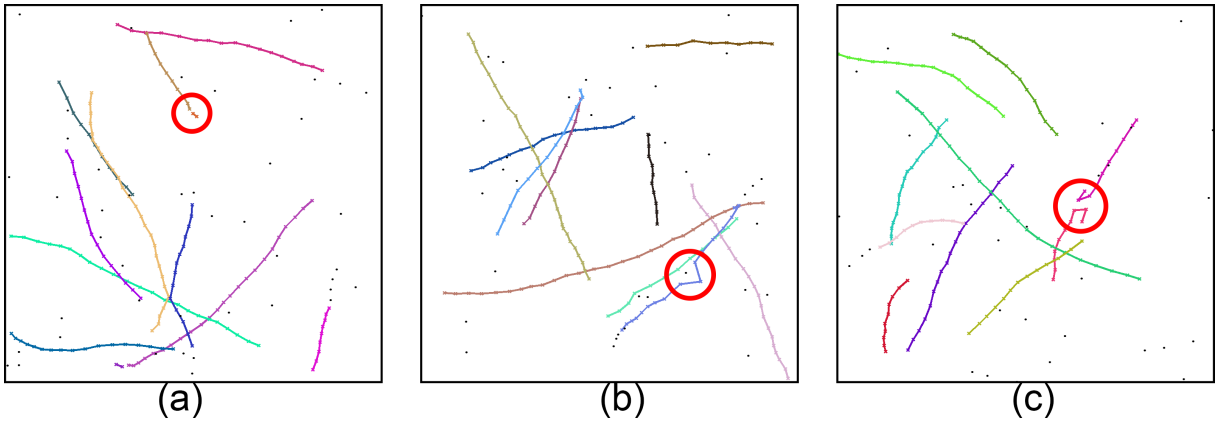



FIGURE 4.4 – Types d'erreurs d'association : (a) fragmentation, (b) association à une fausse alarme, (c) changement d'identité (ID Switch).

**Protocole expérimental et évaluations quantitatives -** Pour chaque valeur de $K$, 5 scenarii de test sont générés, avec une nouvelle partition aléatoire, afin de vérifier la répétabilité des résultats eu égard au caractère stochastique du processus MCMCDA. L'association de données est alors exécutée 10 fois par scénario, totalisant 400 tests. La moyenne des résultats, exprimée à l'aide des mesures introduites précédemment, est ensuite calculée pour chaque valeur de $K$ et illustrée sur la figure 4.5.

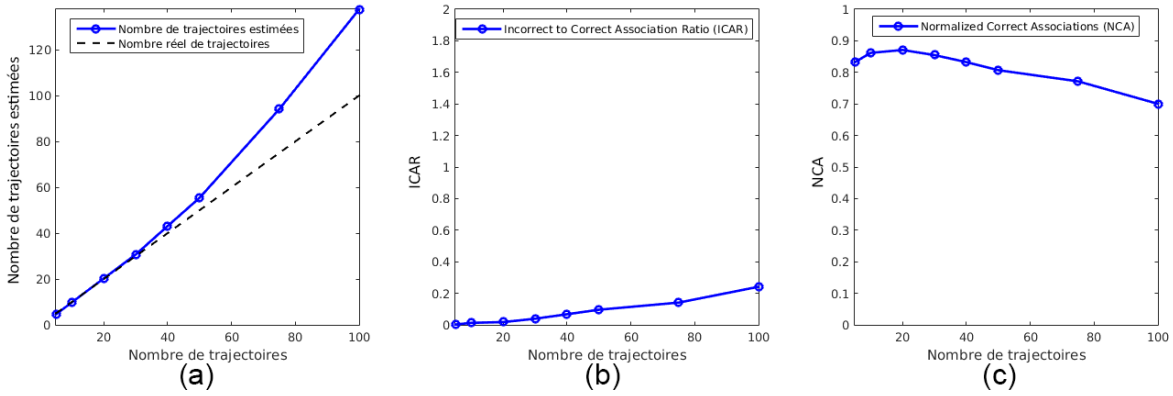



FIGURE 4.5 – Résultats de l'association de données en fonction du nombre de trajectoires et (a) du nombre estimé de trajectoires, (b) du critère ICAR et (c) du critère NCA.

Dans leur ensemble, ces résultats exhibent la grande robustesse de l'approche à des densités élevées de trajectoires. L'écart d'estimation du nombre de trajectoires, en (a) sur la figure 4.5, n'excède pas 11% jusqu'à 50 trajectoires. Pour des valeurs supérieures, nous pouvons observer une divergence atteignant 37% d'écart de trajectoires détectées pour $K = 100$. Elle est principalement due à l'apparition de fragmentations, causées par la non-association de deux observations

successives d'une trajectoire. Le taux d'associations manquées (complémentaire de la métrique NCA en (c) sur la figure 4.5) est ainsi borné à 20% jusqu'à 50 trajectoires et augmente légèrement ensuite jusqu'à 30% pour 100 trajectoires.

Enfin, le taux d'associations incorrectes par rapport aux associations correctes, en (b) sur la figure 4.5, n'excède pas les 10% jusqu'à la limite des 50 trajectoires : en moyenne, pour 10 associations correctes, nous risquons d'obtenir au maximum une association incorrecte.

#### 4.2.3.2 Scénario 2 : variation du taux de fausses alarmes

Ce scénario reprend le précédent mais en fixant le nombre de trajectoires à $K = 10$ et en faisant varier le taux de fausses alarmes $\lambda_f V = [1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$. $\lambda_f V$ représente le nombre moyen de fausses alarmes à chaque instant, sur toute la région $R$. Un exemple du type de partition ainsi créée est illustré sur la figure 4.6.

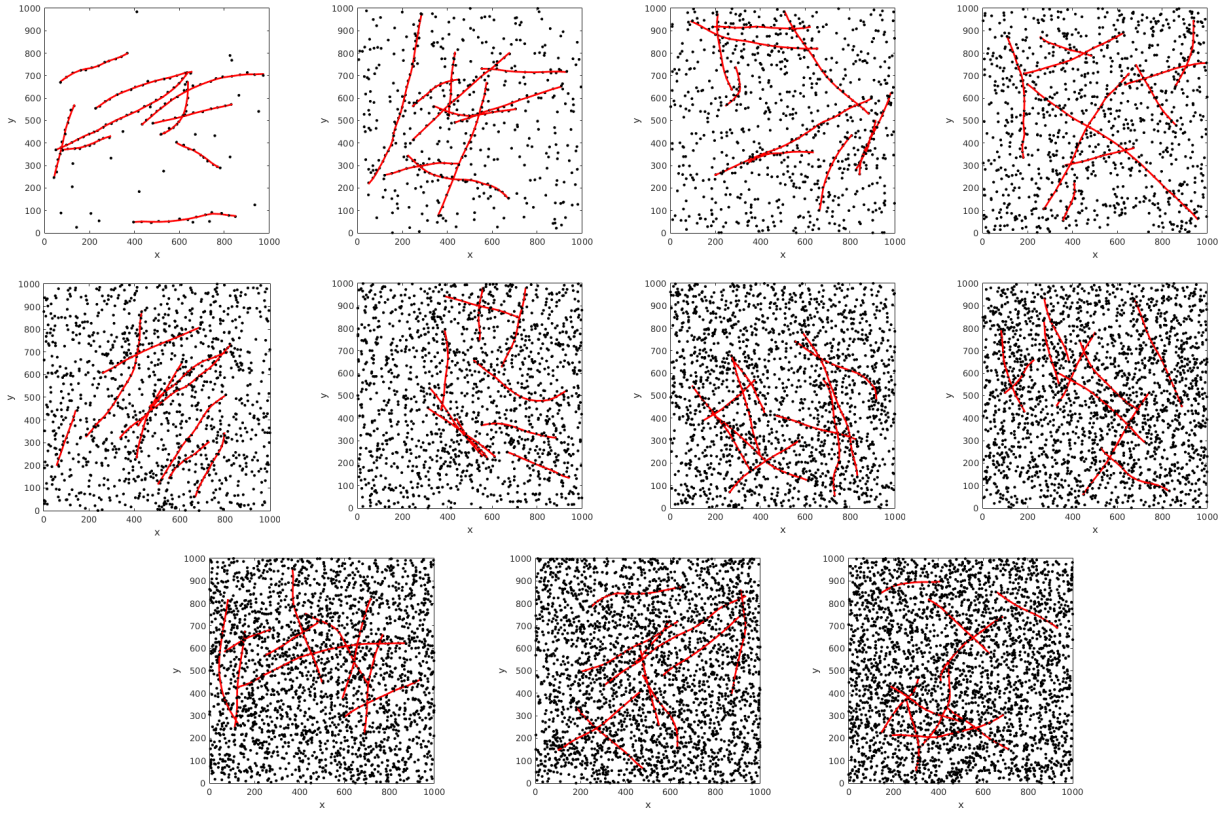


FIGURE 4.6 – De haut-gauche à bas-droit : scenarii de test comportant 10 trajectoires, générées aléatoirement, à plusieurs taux de fausses alarmes par temps et par volume : $\lambda_b V$=[1,10,20,30,40,50,60,70,80,90,100].

**Protocole expérimental et évaluations quantitatives -** Les performances, en terme de nombre de trajectoires, de critères ICAR et NCA, moyennées, sont illustrés sur la figure 4.7.

**Analyse quantitative des résultats -** Malgré un faible nombre de trajectoires, ce scénario est complexe à traiter. Dans le cas le plus extrême, pour $\lambda_b V = 100$, les données sont composées

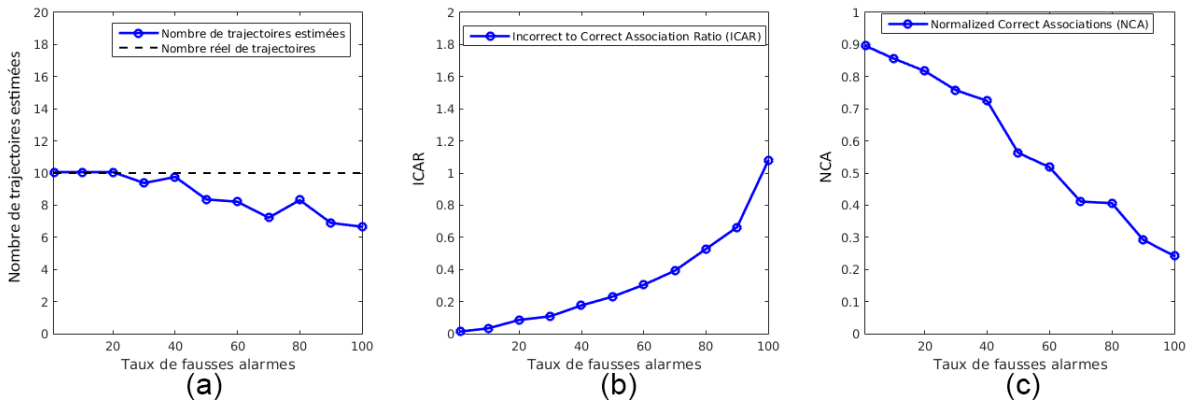



FIGURE 4.7 – Résultats de l'association de données en termes de (a) nombre estimé de trajectoires, (b) critère ICAR et (c) critère NCA.

en moyenne d'une observation issue d'une trajectoire pour 10 fausses alarmes à chaque instant $t$. La figure exhibe une limite qualitative à $\lambda_b V = 40$, en dessous de laquelle l'écart entre le nombre de trajectoires estimées et le nombre réel de trajectoires atteint un maximum de 6,2% (pour $\lambda_b V = 30$). Au delà, cet écart prend des valeurs entre 16% et 33%.

Pour $\lambda_b V \leq 40$, plus de 72% des associations correctes (c) sont réalisées, puis cette valeur chute fortement et à $\lambda_b V = 100$, seul un quart des bonnes observations sont associées. Enfin le taux d'associations incorrectes (b) par rapport aux associations correctes croît exponentiellement mais reste en dessous des 20% à $\lambda_b V = 40$.

#### 4.2.3.3 Scénario 3 : variation de la probabilité de détection

Dans ce scénario, nous relaxons l'hypothèse de persistance des observations, propre aux scénarii 1 et 2. Les observations sont donc détectées avec une probabilité $p_d < 1$ ; on observe donc des non détections, i.e. des faux négatifs. Pour pouvoir traiter ces faux négatifs, nous étendons temporellement la recherche d'observations voisines. Nous fixons à $d_{max} = 3$ le nombre d'observations successives possiblement manquées. Ainsi une observation à $t + 3$ pourra être associée à une observation à $t$ dans un rayon de $d_{max} v_{max} = 150$. Nous fixons également le nombre de trajectoire à $K = 10$, le taux de fausses alarmes à $\lambda_f V = 10$ et nous faisons varier le taux de détection des observations $p_d = [0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$. Un exemple du type de partition ainsi créée est illustré sur la figure 4.8.

**Protocole expérimental et évaluations quantitatives -** Comme précédemment, 5 jeux de données sont créés pour chaque valeur de $p_d$ et traités 10 fois chacun. Les performances moyennées sont illustrées figure 4.9.

**Analyse des résultats -** Comme pour l'analyse des résultats sur les jeux de données des scenarii 1 et 2, il est possible de définir une limite de bon fonctionnement de l'approche. En effet, nous pouvons distinguer qualitativement deux comportements en deçà et au delà de la valeur $p_d = 0.5$. Lorsqu'en moyenne, au moins une observation sur deux est détectée (fausses alarmes exceptées), l'écart d'estimation d'une trajectoire n'excède pas 14%. À l'inverse, pour $p_d = 0.4$ et $p_d = 0.3$, ces écarts atteignent respectivement 35% et 49%.

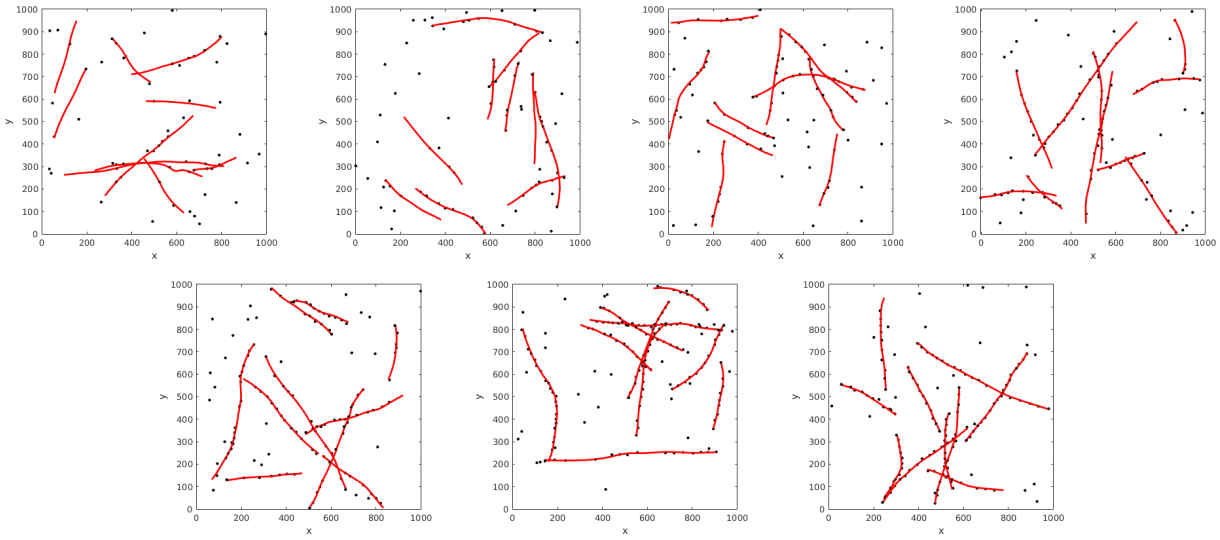


FIGURE 4.8 – De haut-gauche à bas-droit, scenarii de test comportant un taux variable de détection des observations, $p_d = [0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$.

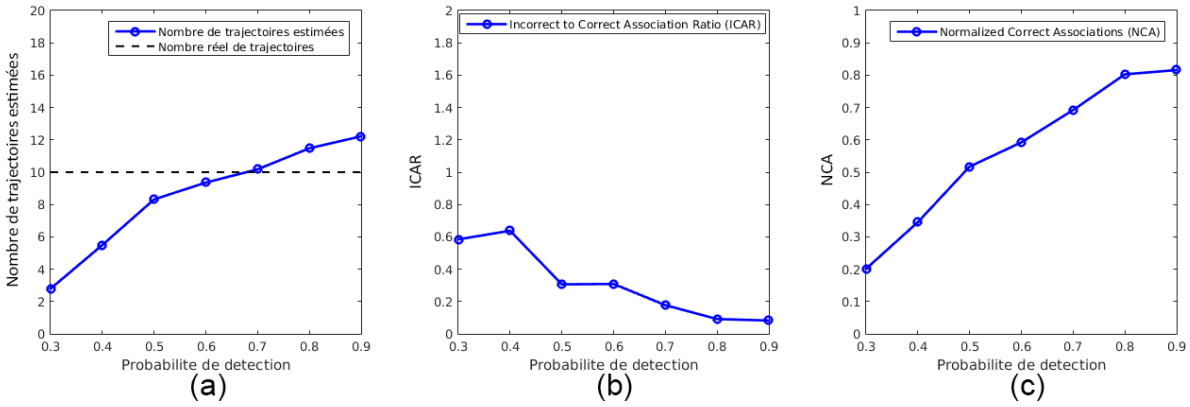



FIGURE 4.9 – Résultats de l'association de données en termes de (a) nombre estimé de trajectoires, (b) ICAR et (c) NCA.

La même rupture est observable sur la mesure du taux d'associations correctes qui atteint quasiment 70% à $p_d = 0.5$ alors que moins de la moitié des associations étaient réalisées à $p_d = 0.4$. De même la mesure de l'ICAR passe en dessous des 15% à $p_d = 0.5$ et double pour $p_d = 0.4$.

**Synthèse des expérimentations -** Pour chacun des trois scenarii, simulant respectivement une grande densité de trajectoires, un taux important de fausses alarmes et un taux de détection faible, des limites de fonctionnement ont pu être fixées, que nous résumons dans le tableau 4.4. Les valeurs indiquées des trois mesures sont les valeurs maximales pour l'erreur d'estimation du nombre de trajectoires (en notant $|\omega|$ le cardinal de la partition $\omega$), les valeurs maximales pour l'ICAR, et les valeurs minimales pour le NCA, toutes dans l'espace borné par les valeurs limites du bon fonctionnement de la méthode.

TABLE 4.4 – Synthèse des résultats sur les 3 jeux de données.

| **Paramètre** | **Valeur limite** | **Erreur de $\|\omega\|$ (%)** | **NCA** | **ICAR** |
|---|---|---|---|---|
| Nombre de trajectoires | 50 | 10.8 | 0.81 | 0.09 |
| Taux de fausses alarmes | 40 | 6.2 | 0.72 | 0.18 |
| Taux de détection | 0.5 | 13.6 | 0.69 | 0.15 |

Suite à ces évaluations synthétiques concluantes, nous décrivons maintenant notre implémentation MCMCDA avec une prise en compte d'observations réalistes audiovisuelles.

## 4.3 Vers le suivi multi-cibles audiovisuel

Notre approche MCMCDA se base précédemment sur un modèle dynamique des cibles à suivre. Ainsi, seules les positions 2D dans le plan du sol des observations sont prises en compte pour les associer ; leur association est donc basée sur la seule cohérence spatiale. Nous intégrons ci-après les signatures audiovisuelles vues aux chapitres 2 et 3.

**Modèle d'apparence -** Un modèle d'apparence est donc considéré pour chaque cible, en plus de son modèle dynamique. Comme évoqué au chapitre 2, ces descripteurs sont issus des régions image englobant les cibles. Il s'agit ici d'histogrammes HSV (pour « *Hue Saturation Value* ») contraints par les axes de symétrie de la cible, comme détaillé en 2.2.2. La contribution majeure de ce chapitre est l'intégration d'un modèle d'apparence, non seulement visuel, mais également sonore, des cibles à suivre, à travers la signature décrite en 2.1.3, sous forme de GMM sur des MFCC. Nous cherchons ainsi à robustifier la fusion des signatures sonores des cibles grâce à l'assistance du traqueur MCMCDA.

### 4.3.1 Intégration des modèles d'apparence

Dans un premier temps, nous nous intéressons à l'intégration de la signature visuelle au sein du MCMCDA. En effet, la position des détections visuelles peut être inférée assez précisément alors que seule une estimation de la distance au microphone peut être extraite à partir du flux audio. L'information apportée par la vidéo est ainsi le socle sur lequel se greffe l'information issue de l'audio. L'intégration des signatures doit ainsi maximiser la vraisemblance des trajectoires portant une grande majorité d'observations appartenant à la même identité et minimiser les autres.

#### 4.3.1.1 Gestion des intermittences des signatures

Une trajectoire est composée d'observations extraites d'instants image non nécessairement consécutifs, en cas de non-détection d'une ou plusieurs observations. Le traitement en logique différée permet de la prolonger jusqu'à une observation future et son état estimé par le filtre de Kalman est alors une suite de prédictions non mises à jour. Dans le cas de suivi audiovisuel nous pouvons alors distinguer 3 cas, en fonction des détections capturées à chaque instant :

— détection audiovisuelle de la cible, sa signature bimodale est donc disponible,
— détection visuelle et non détection sonore de la cible, seule la signature visuelle sera considérée,

— non détection visuelle de la cible avec/sans détection sonore : l'état estimé est prédit à partir de l'état précédent, et seul le modèle dynamique de la cible est employé, éventuellement renforcé par sa signature sonore.

Dans le processus itératif de la méthode MCMC, un ensemble $\omega'$ de trajectoires est proposé et comparé au dernier ensemble accepté $\omega$, et est accepté ou refusé en fonction de sa vraisemblance aux observations $P(\omega'|Y)$ par rapport à celle de la partition courante $P(\omega|Y)$. L'expression de la vraisemblance des modèles vidéo et audio doit ainsi être capable de gérer des comparaisons d'observations hétérogènes et possiblement multimodales (Eq. 4.20). Considérons $\mathbf{y}_t^j$ la j-ième observation au temps $t$ et $\lambda_{AUD}^k$ la signature audio associée à la cible $k$. La log-vraisemblance $\log\left(P\left(\mathbf{y}_t^j|\lambda_{AUD}^k\right)\right)$ de l'appartenance de l'observation $\mathbf{y}_t^j$ en terme de signature audio à la trajectoire $\tau_k$ prenant nécessairement des valeurs négatives, les observations uniquement visuelles seront nécessairement favorisées. De même, les observations non détectées l'emporteront sur une observation détectée et portant une signature visuelle. Afin de compenser les variabilités des dimensions des modèles d'apparence, l'intégration est alors réalisée à travers le rapport de vraisemblance d'une observation à une cible *vs.* un imposteur, ou la différence des log-vraisemblances en échelle logarithmique :

$$\log\left(P\left(\mathbf{y}_t^j|\lambda_{AUD}^k\right)\right) - \log\left(P\left(\mathbf{y}_t^j|\lambda_{AUD}^{UBM}\right)\right) \begin{cases} > 0 \text{ si } \mathbf{y}_t^j \in \tau_k \\ < 0 \text{ si } \mathbf{y}_t^j \notin \tau_k \\ = 0 \text{ en l'absence de signature audio} \end{cases} \tag{4.16}$$

$$\log\left(P\left(\mathbf{y}_t^j|\lambda_{VID}^k\right)\right) - \log\left(P\left(\mathbf{y}_t^j|\lambda_{VID}^{UBM}\right)\right) \begin{cases} > 0 \text{ si } \mathbf{y}_t^j \in \tau_k \\ < 0 \text{ si } \mathbf{y}_t^j \notin \tau_k \\ = 0 \text{ en l'absence de signature vidéo} \end{cases} \tag{4.17}$$

où $\lambda_{AUD}^{UBM}$ est le modèle du monde, détaillé en §2.1.3, modélisant la signature audio d'un locuteur moyen, et par analogie $\lambda_{VID}^{UBM}$ un modèle visuel de l'apparence moyenne d'une personne. La construction de ce dernier modèle est détaillée par la suite.

#### 4.3.1.2 Modèle d'apparence visuel

Les modèles audio et vidéo sont construits de manière similaire. Les scores de rapport de vraisemblance en échelle logarithmique pour l'audio ont été détaillés en 2.1.3, et nous reproduisons la démarche pour l'information vidéo.

Les évaluations de l'association de données sont réalisées en données synthétiques, et nous devons alors construire des modèles d'apparence génériques les plus proches de la réalité. Les données utilisées sont issues des datasets ETHZ [Ess+08], relativement semblables à nos données, et plus fournies. Le corpus, composé d'imagettes de 83 personnes, est scindé en ensemble de personnes cibles (un tiers, soit 28 personnes) et en modèle du monde (deux tiers, soit 55 personnes). Pour chaque cible du corpus d'apprentissage l'ensemble des imagettes correspondantes est lui même divisé en corpus de développement (un tiers) et corpus de test (deux tiers). Les nombres de cibles et d'images utilisés sont résumés dans le tableau 4.5.

Pour chaque individu du modèle du monde, la signature vidéo telle que décrite dans la section 2.2.2 est extraite de chaque image et la moyenne de toutes les signatures (sous forme d'histogrammes HSV) constitue la signature de la cible. La même démarche est réalisée pour l'ensemble d'apprentissage du modèle des cibles. Enfin, les signatures des images de l'ensemble

TABLE 4.5 – Corpus utilisé pour la construction des modèles visuels

| | **Modèle du monde** | **Modèle des cibles** Apprentissage | Test |
|---|---|---|---|
| Nombre de cibles | 55 | 28 | |
| Nombre d'images | 2905 | 1275 | 677 |

de test sont extraites et nous les comparons aux différentes cibles à l'aide de la distance de Bhattacharyya. L'histogramme des distances aux bonnes cibles est illustré en bleu sur la figure 4.10 et l'histogramme des distances au modèle du monde est illustré en rouge.

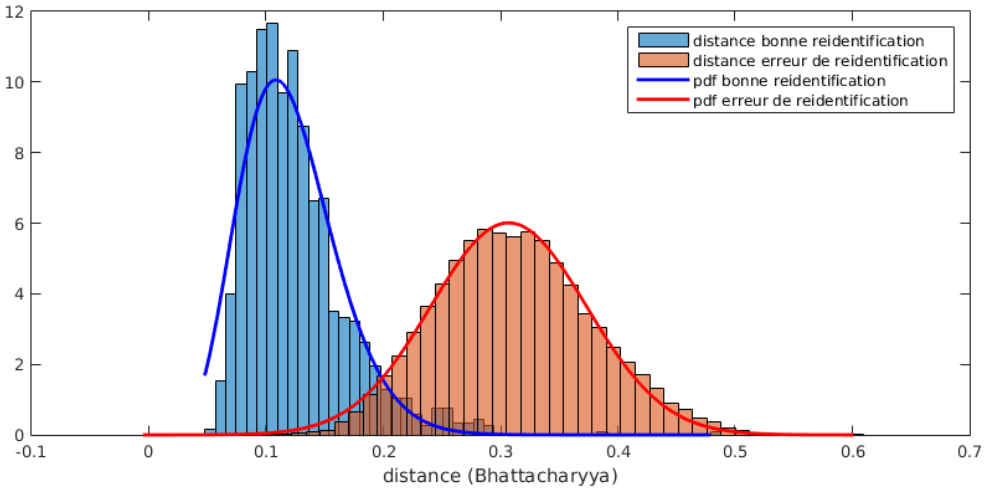



FIGURE 4.10 – Estimation d'un modèle de distance d'une observation à la bonne cible (courbe bleue) et d'un modèle de distance d'une observation à une mauvaise cible (courbe rouge).

Si la distance de la signature d'une image test à une cible du modèle du monde (erreur de réidentification) $\mathrm{d}_{WH}(\mathrm{WH}(I^k_{test}), \mathrm{C}_{UBM})$ (la notation WH se réfère aux histogrammes pondérées *"Weighted Histograms"* par les axes de symétries, qui représente la signature visuelle) a un comportement manifestement gaussien, la distance à la bonne cible $\mathrm{C}_k$ (réidentification correcte) $\mathrm{d}_{WH}(\mathrm{WH}(I^k_{test}), \mathrm{C}_k)$ s'apparie plutôt à une loi Bêta. Nous ajustons alors deux densités de probabilité en conséquence.

$$\mathrm{d}_{WH}(\mathrm{WH}(I^k_{test}), \mathrm{C}_{UBM}) \sim \mathcal{N}(\mu_{UBM}, \sigma^2_{UBM}) \tag{4.18}$$

$$\mathrm{d}_{WH}(\mathrm{WH}(I^k_{test}), \mathrm{C}_k) \sim \mathcal{B}(\alpha, \beta) \tag{4.19}$$

Les paramètres estimés des distributions prennent les valeurs suivantes : $\mu_{UBM} = 0.31$, $\sigma_{UBM} = 0.07$, $\alpha = 8.19$, $\beta = 57.25$. La vraisemblance des observations s'exprime maintenant comme suit :

$$P(\mathbf{Y}|\omega) = \prod_{\tau=\tau_1}^{|\omega|} \prod_{t=1}^{|\tau|} \underbrace{\mathcal{N}(\tau(t)|\mathbf{x}_t(\tau), B_t(\tau))}_{\text{modèle dynamique}} \times \underbrace{\frac{P(\tau(t)|\lambda^{\tau}_{VID})}{P(\tau(t)|\lambda^{UBM}_{VID})}}_{\text{modèle d'apparence}} \tag{4.20}$$

#### 4.3.1.3 Évaluations quantitatives

Afin d'évaluer l'apport de l'intégration d'un modèle d'apparence, les expériences réalisées en section 4.2 sont reproduites en utilisant les mêmes jeux de données synthétisées, pour les trois scenarii comportant respectivement une grande densité de trajectoires, un taux élevé de fausses alarmes et un faible taux de détection. L'évaluation quantitative des performances est effectuée à l'aide des trois métriques précédentes : nombre estimé de trajectoire, taux d'associations incorrectes par rapport aux associations correctes, et taux d'associations correctes réalisées.

Les résultats sont illustrés sur figure 4.11 (a)-(c) pour le scénario faisant varier le nombre de trajectoires, figure 4.11 (d)-(f) pour le scénario faisant varier le taux de fausses alarmes, et figure 4.11 (g)-(i) pour le scénario faisant varier le taux de détection. Les performances pour le MCMCDA avec dynamique seule sont représentées par les courbes bleues et les performances pour le MCMCDA enrichi du modèle d'apparence sont représentés par les courbes rouges.

**Analyse des résultats -** On note d'emblée la réduction sensible du taux d'associations incorrectes dans les trois scenarii ; pour rappel, ceci constituait notre priorité. Le critère ICAR du premier scénario, en figure 4.11 (b), dans la limite de fonctionnement définie ($K = 50$), atteint 0.02 en valeur maximale avec l'intégration du modèle d'apparence, contre presque 0.1 sans. Le gain est encore plus notable dans le deuxième scénario, en figure 4.11 (e), où cette mesure subissait une croissance exponentielle qui disparaît totalement, avec l'ajout du modèle d'apparence : elle est alors bornée à 0.09. Enfin, la même amélioration est visible dans le dernier scénario, en figure 4.11 (h), : l'approche MCMCDA avec apparence conserve un taux d'associations incorrectes bien plus faible qu'à l'aide d'un modèle dynamique seul, notamment à des taux de détection $p_d$ des observations faibles : $\mathrm{ICAR}_{\mathrm{MCMCDA}}(p_d = 0.3) = 0.39$ et $\mathrm{ICAR}_{\mathrm{MCMCDA+Vid}}(p_d = 0.3) = 0.09$.

Ce gain en précision s'accompagne également d'une amélioration du rappel avec l'intégration d'un modèle d'apparence au MCMCDA. Le score NCA, i.e. le taux d'associations correctes réalisées, subit une nette augmentation dans le premier scénario, visible sur la figure 4.11 (c) : pour une densité de trajectoires maximales, 100 trajectoires dans la région $R$, plus de 85% des associations sont réalisées contre 70% en exploitant seulement le modèle dynamique des cibles. Dans la zone de fonctionnement, jusqu'à $K = 50$, ce score s'élève à 90% pour un MCMCDA avec modèle d'apparence, contre 81% sans. Dans le deuxième scénario, en figure 4.11 (f), ce score bénéficie d'un gain d'environ 10% des valeurs minimales dans les limites de la zone de fonctionnement. Ainsi nous avons : $\mathrm{NCA}_{\mathrm{MCMCDA}}(\lambda_b V = 40) = 0.72$ et $\mathrm{NCA}_{\mathrm{MCMCDA+Vid}}(\lambda_b V = 40) = 0.82$. Cependant, pour des valeurs élevées du taux de fausses alarmes, nous observons la même chute de ce score, aussi bien avec des modèle d'apparence, que sans. Dans le dernier scénario, en figure 4.11 (i), l'exploitation d'un modèle d'apparence apporte un gain modeste de 2% à 5% dans la zone de fonctionnement (à partir d'un taux d'observations détectées $p_d = 0.5$).

L'estimation du nombre de trajectoires dans la partition ne tire pas parti en revanche de l'ajout du modèle d'apparence. Dans le premier scénario, en figure 4.11 (a), les deux approches produisent des résultats similaires et dans le deuxième, l'écart au nombre de trajectoires s'est même amplifié, passant de 6.2% maximum dans la zone de fonctionnement à 14.2% en valeur maximale. Enfin, cet écart est légèrement réduit dans le troisième scénario, sur la figure 4.11 (g), où son maximum dans la zone de fonctionnement atteint 7.4% contre 13.6% avec une approche n'exploitant que le modèle dynamique.

**Synthèse :** L'intégration d'un modèle d'apparence au MCMCDA a prouvé sa pertinence en améliorant fortement la précision de sa partition et en proposant des gains intéressants en terme de rappel, dans les trois cas de figure étudiés : variations successives du nombre de trajectoires,

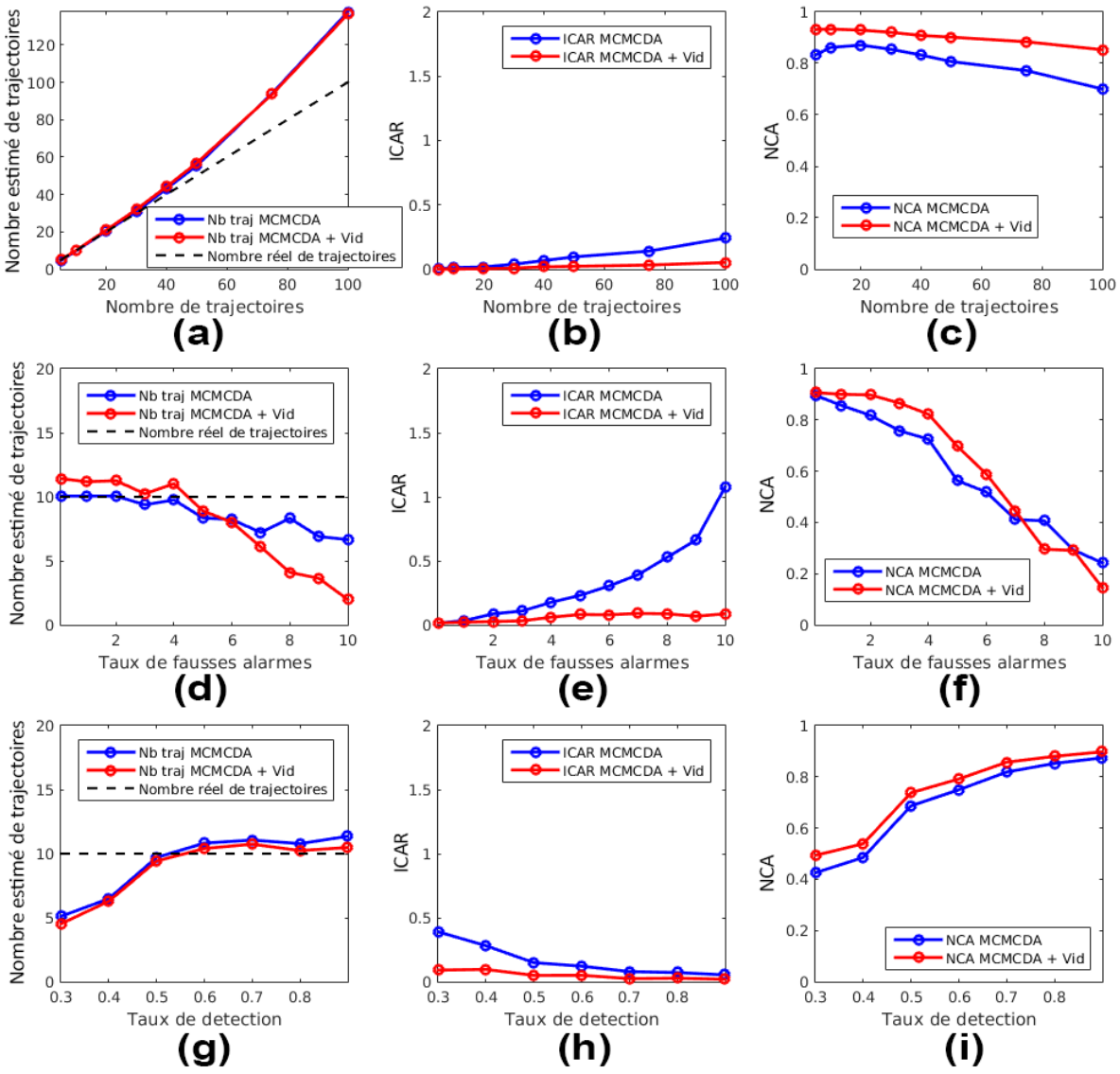



FIGURE 4.11 – Comparaison des résultats de l'approche MCMCDA avec modèle dynamique seul (courbes bleues) et en ajoutant un modèle d'apparence visuel (courbes rouges).

du taux de fausses alarmes et du taux de détection des observations. Les résultats quantitatifs sont synthétisés dans le tableau 4.6. Pour chaque score (noté ./.), la première valeur correspond au score donné par le MCMCDA initial, et la seconde par le MCMCDA avec ajout du modèle d'apparence. En reprenant la nomenclature du tableau 4.4, les valeurs indiquées des trois mesures sont les valeurs maximales pour l'erreur d'estimation du nombre $|\omega|$ de trajectoires et l'ICAR, et les valeurs minimales pour le NCA, toutes dans l'espace borné par les valeurs limites du bon fonctionnement de la méthode.

TABLE 4.6 – Synthèse des résultats sur les 3 jeux de données

| **Paramètre** | **Valeur limite** | **Erreur de $\|\omega\|$ (%)** | **NCA** | **ICAR** |
|---|---|---|---|---|
| Nombre de trajectoires | 50 | **10.8**/13.7 | 0.81/**0.90** | 0.09/**0.02** |
| Taux de fausses alarmes | 40 | **6.2**/14.2 | 0.72/**0.82** | 0.18/**0.06** |
| Taux de détection | 0.5 | 13.6/**7.4** | 0.69/**0.74** | 0.15/**0.05** |

### 4.3.2 Intégration des signatures audio

Les évaluations précédentes ont démontré l'apport global de modèles d'apparence visuelle dans l'algorithme MCMCDA sur les jeux de données relatifs aux trois scénarios. L'étude porte ici sur les gains induits par l'ajout de signatures sonores des cibles. Cette démarche est motivée par le caractère intermittent des deux modalités et de leur possible complémentarité. En effet l'absence de détection vidéo et audio n'auront pas les mêmes causes (e.g. occultation pour la vidéo, interruption du discours pour l'audio) et n'interviendront pas nécessairement aux mêmes instants.

Le challenge du signal audio réside dans l'absence de localisation précise des locuteurs. Comme évoqué au chapitre 3, l'information spatiale disponible est la distance estimée de la source sonore au microphone. Dans le cas mono-cible la fusion audiovisuelle s'appuyait sur la cohérence et la compatibilité spatiale des percepts audio et vidéo à chaque instant $t$. Dans le cas multi-cibles, afin de lever les ambiguïtés des identités des trajectoires la fusion s'appuie sur la compatibilité spatio-temporelle des détections audio et vidéo sur toute la fenêtre d'observation. La fusion est alors séquencée comme suit :

1. Génération de la partition $\omega(i)$ à l'itération $i$ de l'algorithme MCMC,
2. Recherche des associations audiovisuelles candidates à chaque instant $t$ par l'étude de leur cohérence et compatibilité spatiale,
3. Association des identités sonores et visuelles par couple majoritaire.

La gestion des intermittences audio et vidéo a été évoquée en début de section, pour lesquelles 3 cas ont été isolés en fonction de la présence ou l'absence d'information sonore et visuelle. Considérons une trajectoire, représentée sur la figure 4.12 par la ligne continue orange. Elle est initiée au temps $t = 1$ et se poursuit jusqu'à $t = 7$.

La construction de cette trajectoire s'appuiera sur les observations présentes aux temps $t = [1, 2, 3, 6, 7]$, représentées dans les cercles au contour orange. Les observations aux temps $t = 4$ et $t = 5$ ne sont elles pas détectées, ainsi leurs états sont prédits par le filtre de Kalman aux positions représentées par des croix oranges. En parallèle des détections audio sont présentes aux temps $t = [2, 3, 4, 5, 6]$ pour lesquelles la distance au microphone peut être extraite. Les possibles associations audiovisuelles sont représentées par les cercles noirs. Les IDs audiovisuelles $(\mathrm{ID}^{\tau}_{\mathrm{VID}}, \mathrm{ID}^{\tau}_{\mathrm{AUD}})$ peuvent alors être déterminées par occurrence majoritaire respective dans la trajectoire :

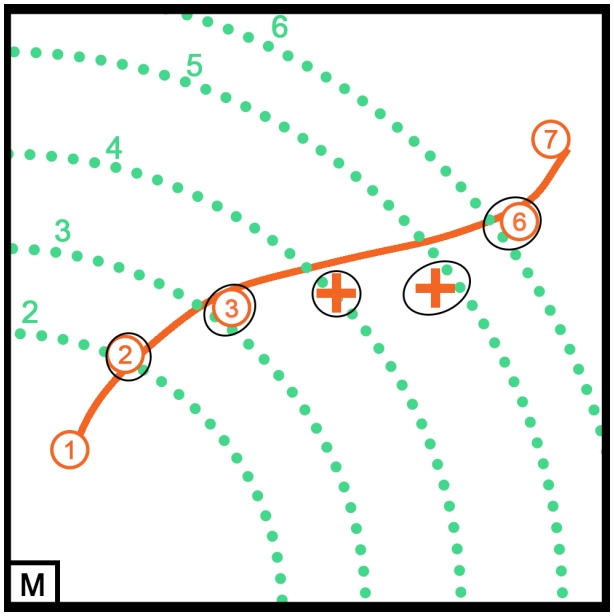



FIGURE 4.12 – Principe d'intégration des signatures audiovisuelles dans le suivi d'une trajectoire

$$\mathrm{ID}^{\tau}_{\mathrm{VID}} = \mathrm{argmax}_i \left( \sum_{\mathrm{id}_{\mathrm{vid}}} \mathrm{id}_{\mathrm{vid}} = i \right) \tag{4.21}$$

$$\mathrm{ID}^{\tau}_{\mathrm{AUD}} = \mathrm{argmax}_i \left( \sum_{\mathrm{id}_{\mathrm{aud}}} \mathrm{id}_{\mathrm{aud}} = i \right) \tag{4.22}$$

Les identités respectivement sonores et visuelles les plus récurrentes sont alors verrouillées et forment la signature audiovisuelle de la cible traquée. La vraisemblance $P(\mathbf{Y}|\omega)$ des observations respectant le couple $(\mathrm{ID}^{\tau}_{\mathrm{VID}}, \mathrm{ID}^{\tau}_{\mathrm{AUD}})$ s'exprime alors en fonction du modèle dynamique des cibles et des signatures visuelle et sonore :

$$P(\mathbf{Y}|\omega) = \prod_{\tau=\tau_1}^{|\omega|} \prod_{t=1}^{|\tau|} \underbrace{\mathcal{N}(\tau(t)|\mathbf{x}_t(\tau), B_t(\tau))}_{\text{modèle dynamique}} \times \underbrace{\frac{P(\tau(t)|\lambda^{\tau}_{VID})}{P(\tau(t)|\lambda^{UBM}_{VID})}}_{\text{modèle visuel}} \times \underbrace{\frac{P(\tau(t)|\lambda^{\tau}_{AUD})}{P(\tau(t)|\lambda^{UBM}_{AUD})}}_{\text{modèle sonore}} \tag{4.23}$$

#### 4.3.2.1 Évaluations du verrouillage des signatures audiovisuelles

Dans le chapitre 3, nous avons introduit une méthode d'association des signatures sonores et visuelles d'un individu par leur cohérence spatiale : à un instant $t$, si la distance euclidienne minimale entre deux détections hétérogènes est inférieure à un seuil de tolérance, les deux signatures sont verrouillées. Cette association est réalisée sous l'hypothèse d'absence d'ambiguïtés entre identités : un unique couple de signatures est présent dans la zone de verrouillage. Si cette hypothèse est valide dans un contexte mono-cible, elle est non applicable en contexte multi-cibles dès lors que des individus se croisent ou sont trop proches les uns des autres.

La mise en place d'un suivi multi-cibles permet alors d'effectuer ce verrouillage de manière isolée, trajectoire par trajectoire, et ainsi de le traiter de manière similaire au cas mono-cible. Le problème s'apparentant à de la classification, nous évaluerons l'association en terme de précision (rapport du nombre de bons couples d'observations créés au nombre total de couples créés) et de rappel (rapport du nombre de bons couples d'observations créés au nombre total de vrais couples possibles). Afin de valider l'apport du suivi, nous comparons la méthode d'association avec celle introduite dans le chapitre 3, qui n'utilise que la cohérence et compatibilité spatiale des observations. Si plusieurs couples sont candidats dans le même voisinage, le choix est tiré aléatoirement parmi eux.

**Évaluations quantitatives de classification -** Le protocole expérimental des sessions précédentes est reproduit à l'identique, mais seul le scénario 3, faisant varier le taux de détection des observations visuelles est évalué. En effet l'information audio propose de renforcer la vraisemblance des trajectoires aux instants de non-détection visuelle. En revanche son apport est négligeable si l'observation est détectée visuellement, l'information vidéo étant localisable bien plus précisément. Nous ne nous intéressons alors qu'à des taux de détections $p_d < 1$. Nous évaluons ci-après le verrouillage des identités visuelles et sonores au niveau de l'ensemble des observations en terme de précision (« combien de verrouillages corrects parmi tous les verrouillages réalisées ? ») et de rappel (« combien de verrouillages corrects parmi tous les verrouillages qu'il est possible de réaliser ? »). Les résultats sont illustrés en figure 4.13.

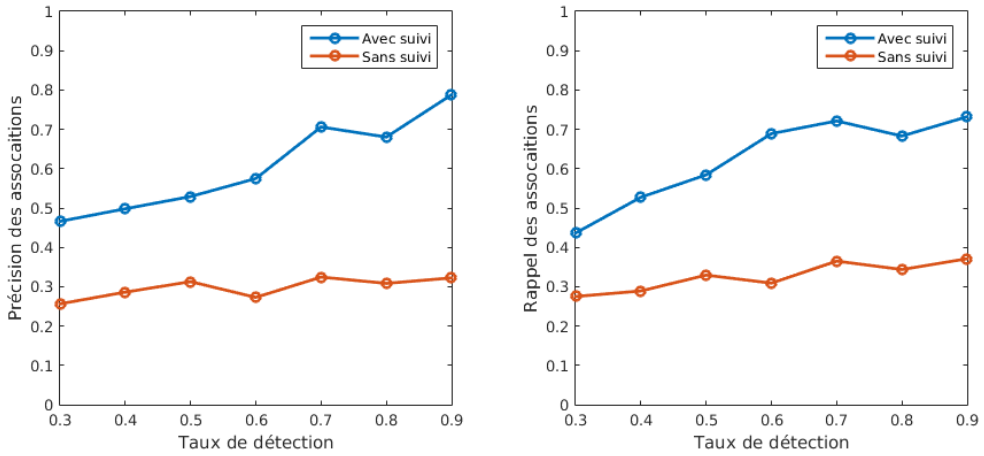



FIGURE 4.13 – Résultats d'associations audiovisuelles, en terme de précision et de rappel, avec/sans suivi multi-cibles.

Les critères rappel et précision oscillent entre 0.3 et 0.4, quel que soit le taux de détection. En revanche, à partir d'un taux de détection de 0.6, environ 70% des associations sont réalisées en s'appuyant sur le suivi multi-cibles. En outre, aux taux de détection les plus faibles, la précision des associations approche 0.5 et atteint 0.8 aux taux les plus élevés, marquant un gain de précision de 0.5 par rapport à la stratégie sans suivi multi-cibles. ces évaluations valident définitivement la plus-value du MCMCDA pour associer les signatures audiovisuelles.

#### 4.3.2.2 Évaluations du suivi multi-cibles audiovisuel

Le verrouillage des signatures audiovisuelles et leur intégration dans le calcul de la vraisemblance de la partition proposée sont réalisées en parallèle, à chaque itération de l'algorithme MCMC. Nous avons démontré ci-dessus l'efficacité du traqueur pour le verrouillage, et nous étudions ici sa plus-value pour la tâche de suivi donc eu égard aux métriques MOT.

Le protocole expérimental décrit en §4.2.3.3, faisant varier le taux de détection des observations, est utilisé pour évaluer l'apport de l'intégration de la signature audiovisuelle de la cible. À titre de comparaison et afin d'identifier les limites de l'approche, les performances du système sont comparées avec celles de l'approche n'utilisant que le modèle dynamique et la signature vidéo, ainsi qu'avec un système simulé avec modèle dynamique, signature vidéo et audio, et dont les positions dans le plan du sol des observations audio $(x_{aud}, y_{aud})$ seraient connues par l'ajout d'un deuxième microphone, par triangulation. Les résultats en terme de MOTA ("Multiple Object Tracking Accuracy") sont illustrés figure 4.14.

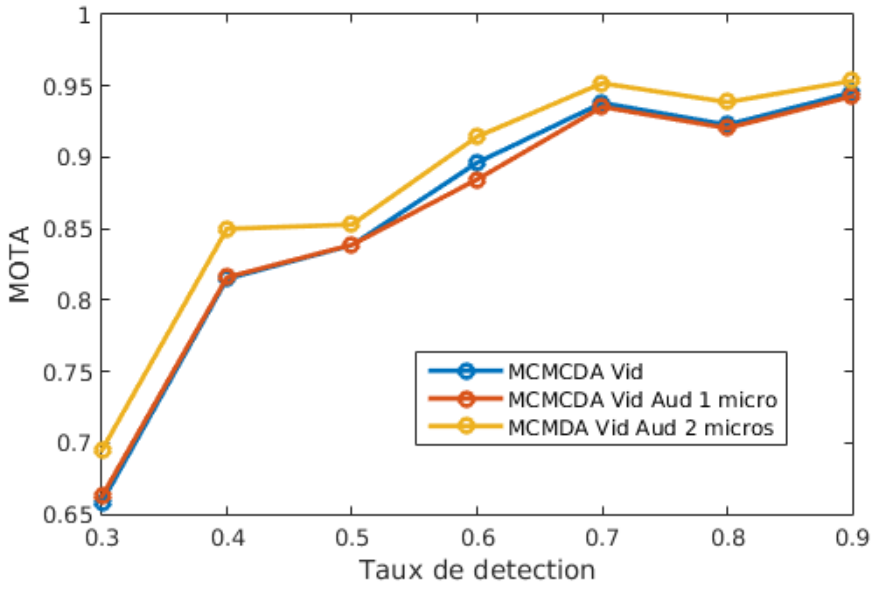



FIGURE 4.14 – Critère MOTA : MCMCDA+signature visuelle avec/sans signature audio (configurations à 1 ou 2 microphones).

Un premier constat face à ces résultats est l'absence de gain apporté par l'information audio, dans la configuration à un microphone, par rapport au MCMCDA avec uniquement la signature vidéo. En revanche, l'ajout d'un second microphone, en contraignant les localisations des observations audio, améliore les performances du traqueur. Nous avons ainsi quantifié la plus-value induite par des détections sonores plus précises.

Concernant l'information visuelle, la signature associée est discriminante sous hypothèse que les apparences vestimentaires des cibles perçues sont différentes. Il nous a semblé opportun d'évaluer en levant cette hypothèse, donc de reproduire l'expérimentation précédente sans signature visuelle. Les résultats sont illustrés en figure 4.15.

Pour une configuration à un microphone, l'ajout de la signature audio dans le MCMCDA est peu probant et illustre ici encore la nécessité d'une meilleure localisation audio pour obtenir des gains significatifs ($> 5\%$). Ces évaluations quantitatives (statistiques) sont complétées ci-après

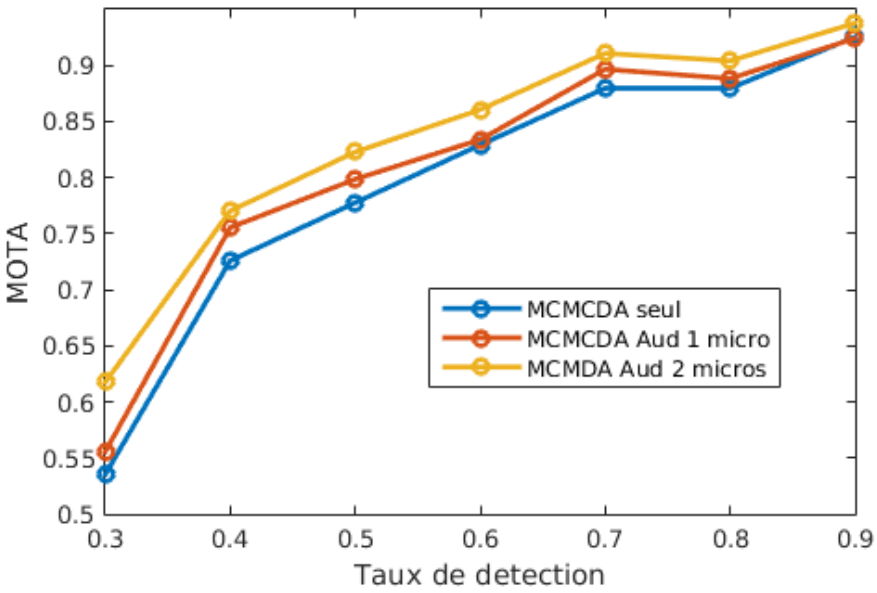



FIGURE 4.15 – Critère MOTA : MCMCDA seul vs. MCMCDA avec signature audio (configurations à 1 et 2 microphones).

par une analyse qualitative donc sur des cas pratiques.

#### 4.3.2.3 Analyse qualitative avec ambiguïtés visuelles

La signature audio permet, dans certaines situations évidentes, de robustifier le maintien des IDs des cibles. Considérons, par exemple, le scénario suivant : deux cibles avec signatures visuelles similaires, qui se rapprochent puis s'éloignent. De manière instinctive, deux hypothèses de partitions peuvent être suggérées : une correcte, maintenant l'identité des deux cibles et une, incorrecte, les intervertissant leurs IDs. Le scénario est illustré en (a) sur la figure 4.16, et les partitions proposées en (b) et (c), respectivement correctes et incorrectes.

Nous évaluons alors les composantes des vraisemblances, exprimées en Eq. 4.23, de ces deux hypothèses de partitions. Celles-ci sont listées dans le tableau 4.7.

TABLE 4.7 – Vraisemblance des partitions avec/sans ID Switch.

| **Log-vraisemblance** | **Partition correcte** | **Partition incorrecte** |
|---|---|---|
| Modèle dynamique | -35.8 | **-28.8** |
| Signature audio | **52.9** | 20.4 |
| Modèle dynamique + audio | **17.1** | -8.4 |

Le modèle dynamique seul dans ce scénario échoue à distinguer la meilleure partition. En effet la partition incorrecte est tout à fait probable d'un point de vue dynamique, et a même une vraisemblance plus élevée que la partition correcte, les cibles effectuant un virage, plus contraignant qu'une trajectoire rectiligne dans un modèle de vitesse constante. L'information

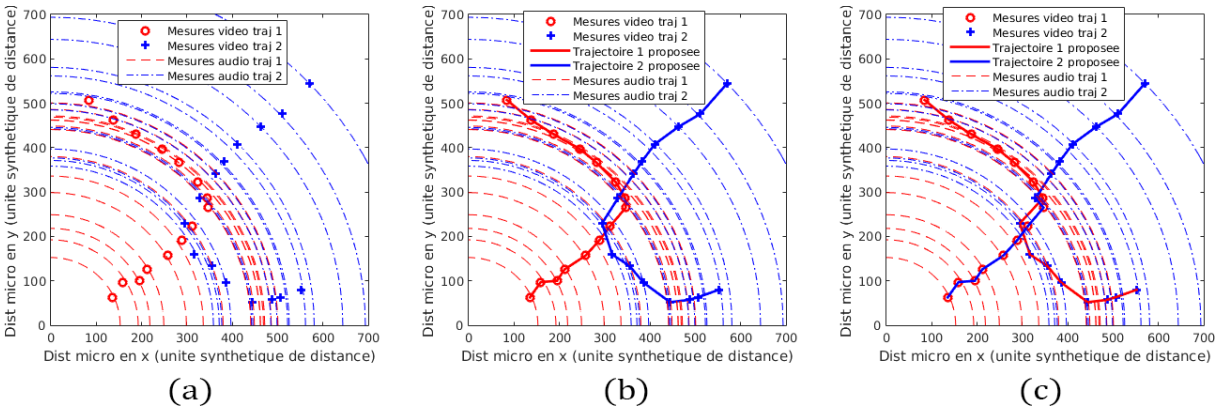



FIGURE 4.16 – Scénario avec deux cibles induisant des observations ambiguës en (a), et deux propositions de partition, correcte en (b) et une incorrecte avec changement ID en (c).

audio contenue dans chaque trajectoire permet en revanche de relever la vraisemblance de la partition correcte et ainsi de lever l'ambiguïté des identités.

En pratique, dans le processus MCMC, ce cas s'apparente à l'hypothèse d'un mouvement d'interversion. L'hypothèse d'interversion de la partition incorrecte à la partition correcte sera alors validée uniquement en tenant compte de l'information apportée par les signatures sonores des cibles, la différence des log-vraisemblances étant positive.

## Conclusion

Les stratégies d'associations audiovisuelles présentées au chapitre 3 sont logiquement inopérantes dans le cas multi-cibles. En effet, elles sont basées sur la seule cohérence spatiale à l'instant courant. En contexte multi-cibles, nous avons proposé un traqueur MOT gérant les signatures sonores et visuelles des cibles.

Inspirée d'une méthode employant des méthodes MCMC pour réaliser l'association de données (MCMCDA) en utilisant un simple modèle dynamique des cibles à suivre, nous avons intégré les signatures des personnes au cœur du processus. L'association de données est réalisée en logique différée, en estimant la vraisemblance d'une partition des observations, soit leur répartition en trajectoires ou en fausses alarmes. Cette vraisemblance est enrichie par les signatures audio et vidéo, et ces dernières sont dans le même temps associées en utilisant leur cohérence et compatibilité spatio-temporelle sur un horizon temporel pour agréger un maximum d'information.

Des évaluations ont été réalisées sur données synthétiques, en simulant plusieurs scenarii, comportant respectivement une grande densité de trajectoires, un taux élevé de fausses alarmes et un taux variable de détection des observations. L'intégration d'une signature visuelle des cibles filtre efficacement un grand nombre de fausses associations, et le mécanisme d'association des signatures audiovisuelles tire clairement parti du traqueur implémenté. En cas de signature visuelle déficiente (vêtements de même couleur), le pouvoir de disambiguation de la signature audio s'observe plutôt via une analyse qualitative et intuitive, certes ébauchée ici, sur des cas particuliers, donc statistiquement peu rencontrées. La plus-value quantitative de la signature audio est notable pour une configuration étendue à plusieurs microphones i.e. pour une localisation plus précise des sources sonores.

Notons enfin que le traqueur MOT améliore ses performances (MOTA) si on intègre notre mécanisme d'association de signature multimodale. En d'autres termes, cette stratégie de verouillage de signatures et le traqueur par MCMCDA se bonifient mutuellement.

# Conclusion

## Synthèse de nos travaux

Les travaux présentés dans ce manuscrit se sont focalisés sur l'apprentissage de signatures audiovisuelles de cibles transitant dans le champ de vue et d'écoute, respectivement de caméras et de microphones ambiants installés de manière éparse. Cette caractérisation non nominative de l'identité sonore et visuelle d'un individu constitue la brique atomique de son activité en terme d'interactions avec ses semblables et avec les infrastructures du bâtiment occupé. Rester à une description bas niveau de l'activité favorise ainsi son intégration dans l'opération neOCampus, à large échelle et pluridisciplinaire, dans laquelle s'inscrivent nos travaux de thèse.

La problématique de cette thèse s'est distinguée par son caractère novateur : les générations dans notre contexte de signatures visuelles d'une part, et sonores d'autre part ne constituent en elles-même pas une contribution, mais le verrouillage, soit l'association correcte des deux modalités correspondant à la même cible, représente un défi important et peu traité dans des contextes de capteurs ambiants et épars. L'absence de corrélation entre les signatures sonores et visuelles a imposé de les associer par localisation des observations sources. Cependant l'instrumentalisation bas coût de la salle d'acquisition ne permet pas d'appliquer les approches classiques de localisation de sources sonores, qui exploitent généralement des indices multi-auraux. Une nouvelle estimation spatiale d'une source sonore en mono-canal a du ainsi être investiguée. L'approche de verrouillage audiovisuel proposée a montré son efficacité dans le cas mono-cible, ou lorsque les cibles sont spatialement très distinctes, mais a atteint ses limites sur des scenarii multi-cibles, donc comportant de nombreuses ambiguïtés d'association audio-vidéo.

Le manuscrit a été structuré en 4 chapitres. Le chapitre 1 a présenté les méthodes de l'état de l'art les plus courantes en reconnaissance de locuteurs et en ré-identification visuelle de personnes, ces tâches étant respectivement responsables de la génération d'une signature sonore et d'une signature visuelle de la cible. Nous avons ainsi justifié le choix des méthodes, outils, bases de données et métriques d'évaluations utilisées. Ceux-ci ont été jugés pertinents eu égard à notre contexte applicatif. Ainsi, la signature sonore d'un locuteur est modélisée par un système GMM-UBM construit sur un vecteur de coefficients cepstraux, architecture traditionnelle en reconnaissance de locuteurs, et adaptée à la nature relativement simple des données traitées. La signature visuelle de la cible s'appuie, quant à elle, sur une accumulation de descripteurs locaux pondérés par les axes de symétrie de la silhouette de la cible, suivant une approche très populaire et performante de la littérature, qui ne nécessite en outre pas de calibration au préalable.

Le chapitre 2 a été consacré à la description et à l'évaluation des éléments constitutifs des chaînes de génération des signatures. En vidéo, la détection visuelle de personnes a été réalisée par l'approche ACF, eu égard à la littérature dédiée et aux évaluations sur des bases de données publiques ; la partie descripteur a été réduite à l'extraction d'histogrammes HSV, sous ensemble de l'approche sélectionnée au chapitre 1, gagnant ainsi en coût calculatoire ; l'appariement a été effectué par la distance de Bhattacharyya. Côté sonore, la détection d'activité vocale a été réalisée

par l'analyse de la modulation de l'énergie à 4 Hertz. Contrairement à la vision, les bases de données couramment utilisées en audio ne correspondait pas à notre contexte applicatif et nous avons alors construit des jeux de données audiovisuels simples et contrôlés (peu de locuteurs, déplacements scénarisés) sur lesquels nous avons validé les différentes briques de notre système.

Le chapitre 3 présente deux stratégies de verrouillage audiovisuel de signatures. La localisation précise des sources n'étant pas réalisable avec l'instrumentation de la pièce, nous avons dans un premier temps extrait un indice de proximité d'une source sonore à un microphone basé sur le taux de signal de parole à la réverbération de la pièce (SRMR). Cet indice, couplé à un indice visuel similaire de proximité, exprimé comme l'inverse de la distance euclidienne entre l'observation et le microphone, permet de définir des zones de verrouillage (« *lock* ») audiovisuel centré sur le microphone. Les identités sonores et visuelles d'un couple d'observations audio et vidéo se situant conjointement dans cette zone pourront alors être fusionnées. Ces zones sont cependant peu larges, mais une installation judicieuse des microphones dans la pièce peut maximiser leur exploitation. La seconde stratégie de fusion s'appuie sur une estimation de la distance source-microphone à l'aide d'une combinaison linéaire du SRMR et de l'énergie du signal. La combinaison idéale de ces paramètres pour l'estimation de la distance est obtenue par l'analyse canonique des corrélations (ACC). Ainsi, une observation visuelle et une observation sonore sont cohérentes et compatibles spatialement si l'estimation respective de la distance au microphone est suffisamment proche : elles peuvent alors être associées.

Le chapitre 4 étend la deuxième stratégie de fusion au cas multi-cibles. Cette stratégie n'est plus applicable directement, le nombre d'ambiguïtés d'association croissant avec le nombre et la densité des cibles. Nous nous sommes alors intéressés au problème du suivi multi-cibles pour assister cette fusion. Nous avons fait le choix d'utiliser une méthode de suivi probabiliste en logique différée, afin de maximiser l'information traitée (la contrainte de temps réel n'étant pas requise). Nous avons ainsi entièrement implémenté l'algorithme MCMCDA qui réalise l'association de données (création et gestion de trajectoires à partir des observations) par processus MCMC. Des partitions aléatoires (ensemble des trajectoires) sont proposées itérativement et sont acceptées ou refusées selon leur vraisemblance aux observations. Cette suite converge alors vers la partition optimale. L'implémentation a été validée sur des données synthétiques simulant des scenarii présentant respectivement un grand nombre de trajectoires, un taux élevé de fausses alarmes, et un taux de détection des observations faible. Nous avons ensuite proposé une intégration des signatures audio et vidéo, en gérant leurs intermittences possibles, dans le calcul de la vraisemblance de la partition. Cette intégration profonde permet alors d'associer les identités à l'échelle de chaque trajectoire, filtrant ainsi les ambiguïtés provoquées par les autres cibles. Cette fusion audiovisuelle assistée par suivi multi-cibles surpasse alors, en terme de précision et de rappel, la méthode de fusion initiale. L'intégration de la signature visuelle dans le suivi a également grandement amélioré ses performances, notamment sur la réduction des associations incorrectes d'observations dans les trajectoires. L'intégration de la signature audio seule offre des gains quantitatifs moins marqués mais peut toutefois se révéler efficace pour lever certaines ambiguïtés visuelles dans des contextes précis.

## Ouverture et perspectives

Ces travaux se situent au croisement de deux communautés Vision et Audio, et donc de nombreuses problématiques apparaissent : ré-identification de personnes, reconnaissance de locuteurs, suivi multi-cibles, localisation de sources sonores. Ce positionnement légèrement en marge de ces thématiques ont conféré aux travaux un caractère exploratoire, et nous avons été

ainsi régulièrement confronté à la difficulté du manque de cadre standard d'évaluation et de données.

### Perspectives à court terme

Afin de valider l'efficacité de nos méthodes nous avons eu recours à nos propres acquisitions pour les chapitres 2 et 3, et à des données purement synthétiques dans le chapitre 4. Si celles-ci s'appuient certes sur des signatures réelles, l'étape suivante est de réaliser une campagne d'acquisition et d'évaluer les performances des différents outils présentés dans cette thèse dans un environnement humain multi-cibles. Nous avons ainsi réalisé des acquisitions audiovisuelles à plus grande échelle, dans le bâtiment ADREAM du LAAS-CNRS, avec les spécificités suivantes :

— Instrumentation :
  - 3 caméras Point Grey Blackfly calibrées, enregistrement à 8 images par secondes,
  - 2 microphones MXL AC-404, enregistrement mono à 16 kHz, 16 bits,
— 3 sessions de 30 minutes,
— 1 « enseignant » par session et entre 6 et 11 « étudiants ».

Chacune des trois sessions simule une situation pédagogique :
— un Cours Magistral (CM),
— des Travaux Dirigés (TD),
— des Travaux Pratiques (TP).

Les comportements des usagers présentent alors de grandes variabilités inter-sessions : en CM, l'activité est réduite à l'enseignant, en TD certaines interactions s'établissent (passage au tableau, échanges en binômes) et en TP tous les usagers interagissent de manière plus chaotique.

Des exemples des images extraites des caméras sont présentés sur le figure 1 : en (a) la caméra 1 dans le scénario CM, en (b) la caméra 2 pendant le scénario TD, en (c) la caméra 3 pendant le scénario TP.

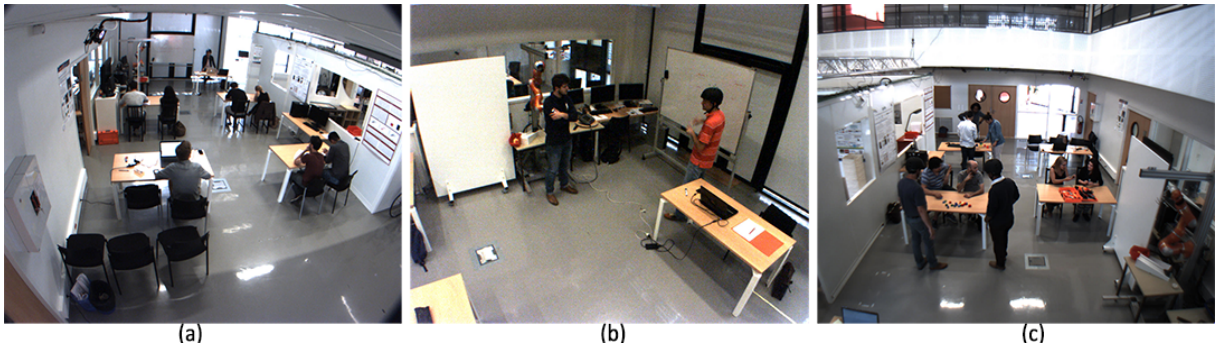



FIGURE 1 – Échantillons visuels du jeu de données en sessions CM (a), TD (b) et TP (c).

Les vidéos ont été annotées manuellement à l'aide de l'outil collaboratif VATIC [15] (« *Video Annotation Tool from Irvine, California* ») qui permet de propager aisément les boîtes englobantes des personnes à travers les trames tout en conservant leur ID. La vérité-terrain audio, en termes de détection d'activité vocale et de locuteurs, a été réalisée à l'aide du logiciel libre

15. `http://carlvondrick.com/vatic/`

Audacity [16]. Nous prévoyons de mettre rapidement ces données à disposition de la communauté scientifique. À notre connaissance, il n'existe pas de bases publiques équivalentes dans la communauté.

### Perspectives à moyen et long termes

Une fois les outils validés sur des données réelles, une interprétation plus haut niveau pourrait être envisagée. Nous nous sommes focalisés dans ces travaux sur des briques atomiques qui définissent les déplacements des usagers d'une pièce et leur signature audiovisuelle. Dans notre contexte pédagogique, ces indices pourraient être transposés à l'identification de concepts plus haut niveau, de la typologie des interactions pour aider à la caractérisation des activités. À titre d'exemple ces interactions pourraient permettre d'identifier le rôle de chaque usager dans la scène (enseignant, étudiant) ou encore la nature du cours dispensé (CM, TD, TP).

Enfin il est possible d'envisager une mise à l'échelle progressive de ces outils. Nous avions émis dans nos travaux l'hypothèse d'un environnement fermé, car nous nous sommes restreints à une salle pédagogique, et à des couples de caméras microphones. La mise à l'échelle peut alors intervenir à deux niveaux : (i) par l'ajout de nouveaux types de capteurs dans le réseau, et l'évaluation de la modularité de ce réseau et de son aspect ≪ *plug and play* ≫, ainsi que (ii) par l'intégration de la topologie du réseau de capteurs lors de la mise à l'échelle de ces outils à l'échelle d'un bâtiment, permettrait de faire transiter les signatures dans le réseau et ainsi étendre la détection de l'activité des usagers dans l'ensemble du bâtiment.

16. `https://audacity.fr/`

# Liste des publications

**Multimedia Tools and Applications** (soumission) Late Fusion Strategies for Characterizing Audiovisual Signatures of Persons *François-Xavier Decroix, Julien Pinquier, Isabelle Ferrané et Frédéric Lerasle*

**10th International Conference on Distributed Smart Camera (ICDSC 2016)** : Online Audiovisual Signature Training for Person Re-identification *François-Xavier Decroix, Julien Pinquier, Isabelle Ferrané et Frédéric Lerasle*

**20ème congrès national sur la Reconnaissance des Formes et l'Intelligence Artificielle (RFIA 2016)** : Apprentissage en ligne d'une signature audiovisuelle pour la réidentification de personnes *François-Xavier Decroix, Frédéric Lerasle, Julien Pinquier et Isabelle Ferrané*

# Table des figures

# Liste des tableaux

# Bibliographie


[Anj+12] A. Anjos et al. “Bob : a free signal processing and machine learning toolbox for researchers”. In : *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan*. Oct. 2012.

[BGS14] Apurva Bedagkar-Gala et Shishir K. Shah. “A survey of approaches and trends in person re-identification”. In : *Image and Vision Computing* 32.4 (2014), p. 270 –286.

[Bha46] A. Bhattacharyya. “On a Measure of Divergence between Two Multinomial Populations”. In : *Sankhyā : The Indian Journal of Statistics (1933-1960)* 7.4 (1946), p. 401–406.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2006.

[BS08] Keni Bernardin et Rainer Stiefelhagen. “Evaluating Multiple Object Tracking Performance : The CLEAR MOT Metrics”. In : *EURASIP Journal on Image and Video Processing* 2008.1 (2008), p. 246309.

[Dec+16] François-Xavier Decroix et al. “Online Audiovisual Signature Training for Person Re-identification”. In : *Proceedings of the 10th International Conference on Distributed Smart Camera*. ICDSC ’16. Paris, France : ACM, 2016, p. 62–68.

[Deh+09] Najim Dehak et al. “Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification”. In : *Interspeech*. Sept. 2009.

[DFP94] Rob Drullman, Joost M. Festen et Reinier Plomp. “Effect of reducing slow temporal modulations on speech reception”. In : *The Journal of the Acoustical Society of America* 95.5 (1994), p. 2670–2680.

[Dik+11] Mert Dikmen et al. “Pedestrian Recognition with a Learned Metric”. In : *Proceedings of the 10th Asian Conference on Computer Vision - Volume Part IV*. ACCV’10. Queenstown, New Zealand : Springer-Verlag, 2011, p. 501–512.

[DLR77] A. P. Dempster, N. M. Laird et D. B. Rubin. “Maximum Likelihood from Incomplete Data via the EM Algorithm”. In : *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), p. 1–38.

[DM80] Steven B. Davis et Paul Mermelstein. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”. In : *ACOUSTICS, SPEECH AND SIGNAL PROCESSING, IEEE TRANSACTIONS ON* (1980), p. 357–366.

[Dol+14] Piotr Dollar et al. “Fast Feature Pyramids for Object Detection”. In : *IEEE Trans. Pattern Anal. Mach. Intell.* 36.8 (août 2014), p. 1532–1545.

[DT05] Navneet Dalal et Bill Triggs. “Histograms of Oriented Gradients for Human Detection”. In : *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*. CVPR '05. Washington, DC, USA : IEEE Computer Society, 2005, p. 886–893.

[ELG07] A. Ess, B. Leibe et L. Van Gool. “Depth and Appearance for Mobile Scene Analysis”. In : *International Conference on Computer Vision (ICCV'07)*. 2007.

[Ess+08] A. Ess et al. “A Mobile Vision System for Robust Multi-Person Tracking”. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*. IEEE Press, 2008.

[Fan60] Gunnar Fant. *Acoustic Theory of Speech Production*. The Hague : Mouton, 1960.

[Far+10] M. Farenzena et al. “Person re-identification by symmetry-driven accumulation of local features”. In : *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2010, p. 2360 –2367.

[FBSS83] T. Fortmann, Y. Bar-Shalom et M. Scheffe. “Sonar tracking of multiple targets using joint probabilistic data association”. In : *IEEE Journal of Oceanic Engineering* 8.3 (1983), p. 173–184.

[Fel+10] P. F. Felzenszwalb et al. “Object Detection with Discriminatively Trained Part-Based Models”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32.9 (2010), p. 1627–1645.

[FF84] K. Fukunaga et T. E. Flick. “An Optimal Global Nearest Neighbor Metric”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.3 (1984), p. 314–318.

[Fis25] R.A. Fisher. *Statistical Methods For Research Workers*. Cosmo study guides. Cosmo Publications, 1925.

[For07] P. E. Forssen. “Maximally Stable Colour Regions for Recognition and Matching”. In : *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, p. 1–8.

[Fur81] S. Furui. “Cepstral analysis technique for automatic speaker verification”. In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.2 (1981), p. 254–272.

[FZC10] T. H. Falk, C. Zheng et W. Y. Chan. “A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech”. In : *IEEE Transactions on Audio, Speech, and Language Processing* 18.7 (2010), p. 1766–1774.

[Gal+05] Sylvain Galliano et al. “The ESTER phase II evaluation campaign for the rich transcription of French broadcast news”. In : *in Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH '05*. 2005, p. 1149–1152.

[Gau+12] N. D. Gaubitch et al. “Performance Comparison of Algorithms for Blind Reverberation Time Estimation from Speech”. In : *IWAENC 2012 ; International Workshop on Acoustic Signal Enhancement*. 2012, p. 1–4.

[GBT07] Doug Gray, Shane Brennan et Hai Tao. “Evaluating appearance models for recognition, reacquisition, and tracking”. In : *In IEEE International Workshop on Performance Evaluation for Tracking and Surveillance, Rio de Janeiro*. 2007.

[GG84] S. Geman et D. Geman. “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), p. 721–741.

[Gil] *Contexts of Accommodation : Developments in Applied Sociolinguistics*. Studies in Emotion and Social Interaction. Cambridge University Press, 1991.

[Gir+13] Ross B. Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In : *CoRR* abs/1311.2524 (2013).

[Gir15] Ross B. Girshick. “Fast R-CNN”. In : *CoRR* abs/1504.08083 (2015).

[GSH06] N. Gheissari, T. B. Sebastian et R. Hartley. “Person Reidentification Using Spatiotemporal Appearance”. In : *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. T. 2. 2006, p. 1528–1535.

[Ham+08] O. Hamdoun et al. “Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences”. In : *2008 Second ACM/IEEE International Conference on Distributed Smart Cameras*. 2008, p. 1–6.

[Has70] W. K. Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In : *Biometrika* 57.1 (1970), p. 97–109.

[Hat+06] Jean-Paul Haton et al. *Reconnaissance Automatique de la Parole Du signal à son interprétation*. UniverSciences (Paris) - ISSN 1635-625X. DUNOD, 2006, p. 392.

[Her90] H. Hermansky. “Perceptual linear predictive (PLP) analysis of speech”. In : *The Journal of the Acoustical Society of America* 87.4 (avr. 1990), p. 1738–1752.

[HM94] H. Hermansky et N. Morgan. “RASTA processing of speech”. In : *IEEE Transactions on Speech and Audio Processing* 2.4 (1994), p. 578–589.

[Hot36] Harold Hotelling. “Relations Between Two Sets of Variates”. In : *Biometrika* 28.3/4 (1936), p. 321–377.

[HS85] T. Houtgast et H. J. M. Steeneken. “A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria”. In : *The Journal of the Acoustical Society of America* 77.3 (1985), p. 1069–1077.

[HSST04] D. R. Hardoon, S. Szedmak et J. Shawe-Taylor. “Canonical Correlation Analysis : An Overview with Application to Learning Methods”. In : *Neural Computation* 16.12 (2004), p. 2639–2664.

[Kal60] R. E. Kalman. “A New Approach to Linear Filtering And Prediction Problems”. In : *ASME Journal of Basic Engineering* (1960).

[Ken+05] Patrick Kenny et al. “Factor analysis simplified”. In : *in ICASSP*. 2005.

[KESM14] E. Khoury, L. El Shafey et S. Marcel. “Spear : An open source toolbox for speaker recognition based on Bob”. In : *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. 2014.

[Kö+12] M. Köstinger et al. “Large scale metric learning from equivalence constraints”. In : *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2012, p. 2288–2295.

[Lam+] Lori F. Lamel et al. “BREF, a Large Vocabulary Spoken Corpus for French”. In : p. 505–508.

[Lar+13] Anthony Larcher et al. “ALIZE 3.0 - open source toolkit for state-of-the-art speaker recognition.” In : *INTERSPEECH*. Sous la dir. de Frédéric Bimbot et al. ISCA, 2013, p. 2768–2772.

[Lee+17] Kong Aik Lee et al. “The I4U Mega Fusion and Collaboration for NIST Speaker Recognition Evaluation 2016”. eng. In : *I4u Mega Fusion and Collaboration for Nist Speaker Recognition Evaluation 2016* (2017).

[Lei+14] Y. Lei et al. “A novel scheme for speaker recognition using a phonetically-aware deep neural network”. In : *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, p. 1695–1699.

[Lis+15] G. Lisanti et al. “Person Re-Identification by Iterative Re-Weighted Sparse Ranking”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.8 (2015), p. 1629–1642.

[LJZ01] Lie Lu, Hao Jiang et HongJiang Zhang. “A Robust Audio Classification and Segmentation Method”. In : *Proceedings of the Ninth ACM International Conference on Multimedia*. MULTIMEDIA ’01. Ottawa, Canada : ACM, 2001, p. 203–211.

[LLM16] A. Larcher, K. A. Lee et S. Meignier. “An extensible speaker identification sidekit in Python”. In : *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, p. 5095–5099.

[LOGP05] Guillaume Lathoud, Jean-Marc Odobez et Daniel Gatica-Perez. “AV16.3 : An Audio-Visual Corpus for Speaker Localization and Tracking”. In : *Machine Learning for Multimodal Interaction : First International Workshop, MLMI 2004, Martigny, Switzerland, June 21-23, 2004, Revised Selected Papers*. Sous la dir. de Samy Bengio et Hervé Bourlard. Berlin, Heidelberg : Springer Berlin Heidelberg, 2005, p. 182–195.

[Mar+15] N. Martinel et al. “Re-Identification in the Function Space of Feature Warps”. In : *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.8 (2015), p. 1656–1669.

[Mar+97] A. Martin et al. “The DET curve in assessment of detection task performance”. In : 1997, p. 1895–1898.

[Met+53] Nicholas Metropolis et al. “Equation of State Calculations by Fast Computing Machines”. In : *The Journal of Chemical Physics* 21.6 (1953), p. 1087–1092.

[Mun57] James Munkres. *ALGORITHMS FOR THE ASSIGNMENT AND TRANSPORTATION PROBLEMS*. 1957.

[NGM01] Elias Nemer, Rafik A. Goubran et Samy A. Mahmoud. “Robust voice activity detection using higher-order statistics in the LPC residual domain”. In : *IEEE Trans. Speech and Audio Processing* 9 (2001), p. 217–231.

[Oh+08] Songhwai Oh et al. *Markov Chain Monte Carlo Data Association for Multi-Target Tracking Univ*. Rapp. tech. 2008.

[PE07] S. J. D. Prince et J. H. Elder. “Probabilistic Linear Discriminant Analysis for Inferences About Identity”. In : *2007 IEEE 11th International Conference on Computer Vision*. 2007, p. 1–8.

[Pov+11] Daniel Povey et al. “The Kaldi Speech Recognition Toolkit”. In : *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Catalog No. : CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society, déc. 2011.

[Pro+10] Bryan Prosser et al. “Person Re-Identification by Support Vector Ranking”. In : *Proc. BMVC*. doi :10.5244/C.24.21. 2010, p. 21.1–11.

[PS01] Jason Pelecanos et Sridha Sridharan. *Feature Warping for Robust Speaker Verification*. 2001.

[PSAo02] Julien Pinquier, Christine Sénac et Régine André-obrecht. “Robust speech / music classification in audio documents”. In : *in Proc. ICSLP'02*. 2002.

[Rat+03] Rama Ratnam et al. “Blind estimation of reverberation time”. In : *The Journal of the Acoustical Society of America* 114.5 (2003), p. 2877.

[RD56] D W Robinson et R S Dadson. “A re-determination of the equal-loudness relations for pure tones”. In : *British Journal of Applied Physics* 7.5 (1956), p. 166.

[Rei79] D. Reid. “An algorithm for tracking multiple targets”. In : *IEEE Transactions on Automatic Control* 24.6 (1979), p. 843–854.

[Ren+15] Shaoqing Ren et al. “Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks”. In : *CoRR* abs/1506.01497 (2015).

[Rey95] Douglas A. Reynolds. “Speaker Identification and Verification Using Gaussian Mixture Speaker Models”. In : *Speech Commun.* 17.1-2 (août 1995), p. 91–108.

[RH89] B. D. Rao et K. V. S. Hari. “Performance analysis of Root-Music”. In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.12 (1989), p. 1939–1949.

[RJ93] Lawrence Rabiner et Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 1993.

[RJO04] R. Ratnam, D. L. Jones et W. D. O'Brien. “Fast algorithms for blind estimation of reverberation time”. In : *IEEE Signal Processing Letters* 11.6 (2004), p. 537–540.

[RPK86] R. Roy, A. Paulraj et T. Kailath. “ESPRIT–A subspace rotation approach to estimation of parameters of cisoids in noise”. In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 34.5 (1986), p. 1340–1342.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri et Robert B. Dunn. “Speaker Verification Using Adapted Gaussian Mixture Models”. In : *Digit. Signal Process.* 10.1 (jan. 2000), p. 19–41.

[Sar+07] M. E. Sargin et al. “Audiovisual Synchronization and Fusion Using Canonical Correlation Analysis”. In : *IEEE Transactions on Multimedia* 9.7 (2007), p. 1396–1403.

[Sat13] Riccardo Satta. “Appearance Descriptors for Person Re-identification : a Comprehensive Review”. In : *CoRR* abs/1307.5748 (2013).

[Sau16] Ferdinand de Saussure. *Cours de linguistique générale*. Paris : Payot, 1916.

[Sch65] M. R. Schroeder. “New Method of Measuring Reverberation Time”. In : *The Journal of the Acoustical Society of America* 37.3 (1965), p. 409–412.

[Sch86] R. Schmidt. “Multiple emitter location and signal parameter estimation”. In : *IEEE Transactions on Antennas and Propagation* 34.3 (1986), p. 276–280.

[SD09a] W. R. Schwartz et L. S. Davis. “Learning Discriminative Appearance-Based Models Using Partial Least Squares”. In : *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. 2009, p. 322–329.

[SD09b] W. R. Schwartz et L. S. Davis. “Learning Discriminative Appearance-Based Models Using Partial Least Squares”. In : *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. 2009, p. 322–329.

[SJJ97] Jeffrey K. Uhlmann Simon J. Julier. *New extension of the Kalman filter to nonlinear systems*. 1997.

[SKS99] Jongseo Sohn, Nam Soo Kim et Wonyong Sung. “A statistical model-based voice activity detection”. In : *IEEE Signal Processing Letters* 6.1 (1999), p. 1–3.

[SL09] Conrad Sanderson et Brian C. Lovell. “Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference”. In : *Advances in Biometrics : Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings*. Sous la dir. de Massimo Tistarelli et Mark S. Nixon. Berlin, Heidelberg : Springer Berlin Heidelberg, 2009, p. 199–208.

[SN09] Ashutosh Saxena et Andrew Y. Ng. “Learning Sound Location from a Single Microphone”. In : *Proceedings of the 2009 IEEE International Conference on Robotics and Automation*. ICRA’09. Kobe, Japan : IEEE Press, 2009, p. 4310–4315.

[SSH13] Seyed Omid Sadjadi, Malcolm Slaney et Larry Heck. *MSR Identity Toolbox v1.0 : A MATLAB Toolbox for Speaker Recognition Research*. Rapp. tech. 2013.

[TGY16] S. Tong, H. Gu et K. Yu. “A comparative study of robustness of deep learning approaches for VAD”. In : *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, p. 5695–5699.

[UN14] UN. “World urbanization prospects”. In : (2014).

[VL98] Olli Viikki et Kari Laurila. “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition”. In : *Speech Commun.* 25.1-3 (août 1998), p. 133–147.

[WHN08] J. Y. C. Wen, E. A. P. Habets et P. A. Naylor. “Blind estimation of reverberation time based on the distribution of signal decay rates”. In : *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2008, p. 329–332.

[WS09] Kilian Q. Weinberger et Lawrence K. Saul. “Distance Metric Learning for Large Margin Nearest Neighbor Classification”. In : *J. Mach. Learn. Res.* 10 (juin 2009), p. 207–244.

[WZ11] J. Wu et X. L. Zhang. “Efficient Multiple Kernel Support Vector Machine Based Voice Activity Detection”. In : *IEEE Signal Processing Letters* 18.8 (2011), p. 466–469.

[ZGX11] W. S. Zheng, S. Gong et T. Xiang. “Person re-identification by probabilistic relative distance comparison”. In : *CVPR 2011*. 2011, p. 649–656.

[Zha+16] S. Zhang et al. “How Far are We from Solving Pedestrian Detection ?” In : *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, p. 1259–1267.

[Ziv04] Z. Zivkovic. “Improved adaptive Gaussian mixture model for background subtraction”. In : *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.* T. 2. 2004, 28–31 Vol.2.

[ZYH16] Liang Zheng, Yi Yang et Alexander G. Hauptmann. “Person Re-identification : Past, Present and Future”. In : *CoRR* abs/1610.02984 (2016).